Consistent Learning by Composite Proximal Thresholding *

Patrick L. Combettes[1], Saverio Salzo[2], and Silvia Villa[3]

[1]Sorbonne Universités – UPMC Univ. Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions
F-75005 Paris, France
plc@ljll.math.upmc.fr

[2]Università degli Studi di Genova, Dipartimento di Matematica, 16146 Genova, Italy
saverio.salzo@unige.it

[3]Massachusetts Institute of Technology and Istituto Italiano di Tecnologia
Laboratory for Computational and Statistical Learning, Cambridge, MA 02139, USA
silvia.villa@iit.it

December 3, 2015

Abstract

We investigate the modeling and the numerical solution of machine learning problems with prediction functions which are linear combinations of elements of a possibly infinite-dimensional dictionary. We propose a novel flexible composite regularization model, which makes it possible to incorporate various priors on the coefficients of the prediction function, including sparsity and hard constraints. We show that the estimators obtained by minimizing the regularized empirical risk are consistent in a statistical sense, and we design an error-tolerant composite proximal thresholding algorithm for computing such estimators. New results on the asymptotic behavior of the proximal forward-backward splitting method are derived and exploited to establish the convergence properties of the proposed algorithm. In particular, our method features a $o(1/m)$ convergence rate in objective values.

1 Introduction

A central task in data science is to extract information from collected observations. Optimization procedures play a central role in the modeling and the numerical solution of data-driven information extraction problems. In the present paper, we consider the problem of learning from examples within the framework of generalized linear models [5, 19, 21]. The goal is to estimate a functional relation f from an input set $\mathcal{X}$ into an output set $\mathcal{Y} \subset \mathbb{R}$. The data set consists of the observation of a finite number of realizations $z_n = (x_i, y_i)_{1 \leqslant i \leqslant n}$ in $\mathcal{X} \times \mathcal{Y}$ of independent input/ouput random pairs with

*The work of P. L. Combettes was supported by the CNRS MASTODONS project under grant 2013MesureHD and by the CNRS Imag'in project under grant 2015OPTIMISME.

an unknown common distribution P. We adopt a generalized linear model, i.e., we assume that the target function f can be approximated by estimators of the form

$$f_u = \sum_{k\in\mathbb{K}} \mu_k \phi_k, \tag{1.1}$$

where $\mathbb{K}$ is at most countable, $u = (\mu_k)_{k\in\mathbb{K}} \in \ell^2(\mathbb{K})$, and $(\phi_k)_{k\in\mathbb{K}}$ is a family of bounded measurable functions from $\mathcal{X}$ to $\mathbb{R}$; such a family is called a *dictionary*, and its elements are called *features*. The estimator $f_{\widehat{u}_{n,\lambda}}$ is computed via the approximate minimization of the convex regularized empirical risk

$$\widehat{u}_{n,\lambda} \in \operatorname{Argmin}^{\varepsilon_n}_{u\in\ell^2(\mathbb{K})} \left(\frac{1}{n}\sum_{i=1}^{n} |f_u(x_i) - y_i|^2 + \lambda \sum_{k\in\mathbb{K}} g_k(\mu_k) \right), \tag{1.2}$$

where $\lambda \in \mathbb{R}_{++}$ and where the convex regularization functions $(g_k)_{k\in\mathbb{K}}$ enforce or promote prior knowledge on the coefficients $(\mu_k)_{k\in\mathbb{K}}$ of the decomposition of the target function f with respect to the dictionary. Our objective is to select a family of regularizers $(g_k)_{k\in\mathbb{K}}$ that model a broad range of prior knowledge and, at the same time, lead to implementable solution algorithms that produce consistent estimators as the sample size n becomes arbitrarily large. To satisfy this dual objective, we shall focus our attention on the following flexible composite model: each function $g_k\colon \mathbb{R} \to \left]-\infty, +\infty\right]$ is of the form

$$g_k = \iota_{C_k} + \sigma_{D_k} + h_k, \quad h_k - \eta|\cdot|^r \in \Gamma_0^+(\mathbb{R}), \quad r \in \left]1,2\right], \quad \eta \in \mathbb{R}_{++}, \tag{1.3}$$

where ι_{C_k} is the indicator function of a closed interval $C_k \subset \mathbb{R}$, σ_{D_k} is the support function of an interval $D_k \subset \mathbb{R}$, $\eta \in \mathbb{R}_{++}$, and $h_k\colon \mathbb{R} \to \mathbb{R}_+$ is convex and such that $h_k(0) = 0$. In this model, the role of C_k is to explicitly enforce hard constraints and the role of D_k is to promote sparsity [9]. On the other hand, h_k provides stability and will be seen to be instrumental in guaranteeing consistency. Note that the model (1.2)–(1.3) refines that considered in [9] and that it encompasses ridge regression [21, 22], elastic net [16, 34], bridge regression [20], and generalized Gaussian models [1]. Proximal thresholders [9], which extend the basic notion of a soft thresholder, will play a key role in our analysis.

The main objective of our paper is to investigate statistical and algorithmic aspects of the estimators based on (1.2)–(1.3). Our main contributions are the following:

- We prove the consistency of the estimators $(f_{\widehat{u}_{n,\lambda}})_{n\in\mathbb{N}}$ as $n \to +\infty$, as well as the convergence of the corresponding coefficients $(\widehat{u}_{n,\lambda})_{n\in\mathbb{N}}$ in $\ell^r(\mathbb{K})$. This generalizes in particular the analysis of [16], which corresponds to the special case when $C_k = \mathbb{R}$, $D_k = [-\omega_k, \omega_k]$, and $h_k = \eta|\cdot|^2$. In this case, (1.3) reduces to

$$g_k = \omega_k|\cdot| + \eta|\cdot|^2. \tag{1.4}$$

- We establish new asymptotic properties for an error-tolerant forward-backward splitting algorithm based on proximal thresholders. In particular, we establish new minimizing properties and a rate of convergence $o(1/m)$ for the objective function values in the presence of variable proximal parameters, relaxations, and computational errors. These results, which are of interest in their own right, improve on the state of the art, which considers either the error free-case and the non-relaxed version [4, 15], or convergence only in an ergodic sense [28].

The paper is organized as follows. In Section 2, we set the problem formally and present the main results concerning the statistical and algorithmic issues pertaining to the proposed estimators. Section 3 is devoted to proving the consistency of the estimators, which is established in Theorem 2.4. In Section 4, we establish Theorem 2.7, which concerns the asymptotic behavior of a proximal forward-backward splitting algorithm, and Theorem 2.11, which specifically deals with the structure considered in (1.2)–(1.3). Additional properties of the regularizers defined in (1.2) are studied in Appendices A and B.

Notation. $\mathbb{N}^* = \mathbb{N} \smallsetminus \{0\}$, $\mathbb{R}_+ = [0,+\infty[$, and $\mathbb{R}_{++} =]0,+\infty[$. Throughout, $\mathbb{K}$ is an at most countably infinite index set. We denote by $(e_k)_{k\in\mathbb{K}}$ the canonical orthonormal basis of $\ell^2(\mathbb{K})$. The canonical norm of $\ell^r(\mathbb{K})$ is denoted by $\|\cdot\|_r$. Let $\mathcal{H}$ be a real Hilbert space. We denote by $\langle\cdot \mid \cdot\rangle$ and $\|\cdot\|$ the scalar product and the associated norm of $\mathcal{H}$. The set of proper lower semicontinuous convex functions from $\mathcal{H}$ to $]-\infty,+\infty]$ is denoted by $\Gamma_0(\mathcal{H})$, and the subset of $\Gamma_0(\mathcal{H})$ of functions valued in $[0,+\infty]$ by $\Gamma_0^+(\mathcal{H})$. Let $\varphi \in \Gamma_0(\mathcal{H})$. The subdifferential of φ at $u \in \mathcal{H}$ is $\partial\varphi(u) = \big\{u^* \in \mathcal{H} \;\big|\; (\forall v \in \mathcal{H})\; \varphi(u) + \langle v-u \mid u^*\rangle \leqslant \varphi(v)\big\}$ and, for every $\varepsilon \in \mathbb{R}_{++}$, $\mathrm{Argmin}^{\varepsilon}_{\mathcal{H}}\varphi = \big\{u \in \mathcal{H} \;\big|\; \varphi(u) \leqslant \inf\varphi(\mathcal{H}) + \varepsilon\big\}$. Let $\mathcal{D} \subset \mathcal{H}$. The indicator function of $\mathcal{D}$ is denoted by $\iota_{\mathcal{D}}$ and the support function of $\mathcal{D}$ is $\sigma_{\mathcal{D}}\colon \mathcal{H} \to]-\infty,+\infty] : u \mapsto \sup_{v\in\mathcal{D}}\langle v \mid u\rangle$. Let $u \in \mathcal{H}$. Then $\mathrm{prox}_{\varphi} u = \mathrm{argmin}_{v\in\mathcal{H}}(\varphi(v) + (1/2)\|u-v\|^2)$ [24]. Suppose that $\mathcal{D}$ is a nonempty, closed, and convex subset of $\mathcal{H}$. Then $\mathrm{prox}_{\iota_{\mathcal{D}}} = \mathrm{proj}_{\mathcal{D}}$ is the projection operator onto $\mathcal{D}$, and $\mathrm{prox}_{\sigma_{\mathcal{D}}} = \mathrm{Id} - \mathrm{proj}_{\mathcal{D}} = \mathrm{soft}_{\mathcal{D}}$ is the soft-thresholder with respect to $\mathcal{D}$. For background on convex analysis and optimization, see [3].

2 Problem setting and main results

The following assumption will be made in our main results.

Assumption 2.1 $(\mathcal{X}, \mathfrak{A}_{\mathcal{X}})$ is a measurable space, $\mathcal{Y} \subset \mathbb{R}$ is a nonempty bounded interval, and $b = \sup_{y\in\mathcal{Y}}|y|$. Moreover, P is a probability measure on $\mathcal{X}\times\mathcal{Y}$ with marginal $P_{\mathcal{X}}$ on $\mathcal{X}$. The *risk* is

$$R\colon L^2(P_{\mathcal{X}}) \to \mathbb{R}_+ : f \mapsto \int_{\mathcal{X}\times\mathcal{Y}} |f(x)-y|^2\, \mathrm{d}P(x,y) \tag{2.1}$$

and $(\phi_k)_{k\in\mathbb{K}}$ is a family of measurable functions from $\mathcal{X}$ to $\mathbb{R}$ such that, for some $\kappa \in \mathbb{R}_{++}$,

$$\sup_{x\in\mathcal{X}} \sum_{k\in\mathbb{K}} |\phi_k(x)|^2 \leqslant \kappa^2. \tag{2.2}$$

The *feature map* is

$$\Phi\colon \mathcal{X} \to \ell^2(\mathbb{K}) \colon x \mapsto (\phi_k(x))_{k\in\mathbb{K}} \tag{2.3}$$

and

$$A\colon \ell^2(\mathbb{K}) \to \mathbb{R}^{\mathcal{X}} \colon u = (\mu_k)_{k\in\mathbb{K}} \mapsto f_u = \sum_{k\in\mathbb{K}} \mu_k\phi_k \;\; \text{(pointwise)}. \tag{2.4}$$

In addition,

(a) $(C_k)_{k\in\mathbb{K}}$ is a family of closed intervals in $\mathbb{R}$ such that $0 \in \bigcap_{k\in\mathbb{K}} C_k$.

(b) $(D_k)_{k\in\mathbb{K}}$ is a family of nonempty closed bounded intervals in $\mathbb{R}$ such that $\sum_{k\in\mathbb{K}} |(\inf D_k)_+|^{r^*} < +\infty$ and $\sum_{k\in\mathbb{K}} |(\inf D_k)_-|^{r^*} < +\infty$.

(c) $(h_k)_{k\in\mathbb{K}}$ is a family in $\Gamma_0^+(\mathbb{R})$ such that $(\forall k \in \mathbb{K})$ $h_k(0) = 0$ and $h_k - \eta|\cdot|^r \in \Gamma_0^+(\mathbb{R})$ for some $r \in \left]1, 2\right]$ and $\eta \in \mathbb{R}_{++}$.

We define

$$\begin{cases} (\forall k \in \mathbb{K}) \quad g_k = \iota_{C_k} + \sigma_{D_k} + h_k \\ F = R \circ A \colon \ell^2(\mathbb{K}) \to \mathbb{R} \\ G\colon \ell^2(\mathbb{K}) \to \left]-\infty, +\infty\right] : u \mapsto \sum_{k\in\mathbb{K}} g_k(\mu_k) \\ \mathcal{C} = \overline{A\big(\ell^2(\mathbb{K}) \cap \bigtimes_{k\in\mathbb{K}} C_k\big)} \quad \text{(closure is taken in } L^2(P_{\mathcal{X}})\text{)}. \end{cases} \tag{2.5}$$

$(X_i, Y_i)_{i\in\mathbb{N}}$ is a sequence of i.i.d. random variables, on an underlying probability space $(\Omega, \mathfrak{A}, \mathsf{P})$, taking values in $\mathcal{X} \times \mathcal{Y}$ and distributed according to P. For every $n \in \mathbb{N}^*$, $Z_n = (X_i, Y_i)_{1\leqslant i\leqslant n}$. The function $\varepsilon\colon \mathbb{R}_{++} \to [0, 1]$ satisfies $\varepsilon(\lambda) \to 0$ as $\lambda \to 0^+$. Moreover, for every $n \in \mathbb{N}^*$, every $\lambda \in \mathbb{R}_{++}$, and every training set $z_n = (x_i, y_i)_{1\leqslant i\leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$

$$\widehat{u}_{n,\lambda}(z_n) \in \operatorname{Argmin}^{\varepsilon(\lambda)}_{u\in\ell^2(\mathbb{K})} \left(\frac{1}{n} \sum_{i=1}^{n} |f_u(x_i) - y_i|^2 + \lambda G(u) \right). \tag{2.6}$$

Remark 2.2

(i) The proposed learning method falls into the class of regularized empirical risk minimization algorithms. However, it differs from the classical setting which uses the squared norm as a regularizer [13, 18, 19].

(ii) The conditions on the sequences $((\inf D_k)_+)_{k\in\mathbb{K}}$ and $((\sup D_k)_-)_{k\in\mathbb{K}}$ given in Assumption 2.1 ensure that $G \in \Gamma_0(\ell^2(\mathbb{K}))$. Moreover, $\operatorname{dom} G \subset \ell^r(\mathbb{K})$ and G is bounded from below and coercive (see Lemma A.1).

(iii) It follows from (2.2) that the linear operator A is well defined and continuous with respect to the topology of the pointwise convergence on $\mathbb{R}^{\mathcal{X}}$, that $\operatorname{ran} A \subset L^\infty(P_{\mathcal{X}})$, and that $A\colon \ell^2(\mathbb{K}) \to L^2(P_{\mathcal{X}})$ is a bounded linear operator such that $\|A\| \leqslant \kappa$. The feature map Φ and A are connected via the identities

$$(\forall k \in \mathbb{K})(\forall x \in \mathcal{X}) \quad \langle \Phi(x) \mid e_k \rangle = (Ae_k)(x). \tag{2.7}$$

In [16, Proposition 3] it is shown that $\operatorname{ran} A$ can be endowed with a reproducing kernel Hilbert space structure for which A becomes a partial isometry, and the corresponding reproducing kernel is

$$K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}\colon (x, x') \mapsto \sum_{k\in\mathbb{K}} \phi_k(x)\phi_k(x'). \tag{2.8}$$

In the above setting, the goal is to minimize the risk R of (2.1) on the closed convex subset $\mathcal{C}$ of $L^2(P_{\mathcal{X}})$ using the n i.i.d. observations $Z_n = (X_i, Y_i)_{1\leqslant i\leqslant n}$. In this respect, recall that the regression function $f^\dagger$ is the minimizer of the risk on $L^2(P_{\mathcal{X}})$ and that

$$(\forall f \in L^2(P_{\mathcal{X}})) \quad R(f) - \inf\, R(L^2(P_{\mathcal{X}})) = \|f - f^\dagger\|^2_{L^2}. \tag{2.9}$$

This means that minimizing R on $L^2(P_{\mathcal{X}})$ is equivalent to approximating the regression function $f^\dagger$. In our constrained setting, the solution to the regression problem on $\mathcal{C}$ results in a target function $f_{\mathcal{C}}$ with the following properties.

Proposition 2.3 *Suppose that Assumption 2.1 is in force. Then there exists a unique $f_{\mathcal{C}} \in \mathcal{C}$ such that $R(f_{\mathcal{C}}) = \inf R(\mathcal{C})$. Moreover, the following hold:*

(i) *$f_{\mathcal{C}}$ is the projection of $f^\dagger$ onto $\mathcal{C}$ in $L^2(P_{\mathcal{X}})$.*

(ii) $(\forall f \in \mathcal{C})\ \|f - f_{\mathcal{C}}\|_{L^2}^2 \leqslant R(f) - \inf R(\mathcal{C})$.

(iii) $$(\forall f \in \mathcal{C})\ \ R(f) - \inf R(\mathcal{C}) \leqslant 2\Big[\Big(\|f - f_{\mathcal{C}}\|_{L^2} + \sqrt{\inf R(\mathcal{C}) - \inf R(L^2(P_{\mathcal{X}}))}\Big)^2 + \inf R(L^2(P_{\mathcal{X}}))\Big]^{1/2}\|f - f_{\mathcal{C}}\|_{L^2}.$$

Proposition 2.3 states that, as in the unconstrained case, minimizing the risk over $\mathcal{C}$ is still equivalent to approaching $f_{\mathcal{C}}$ in $L^2(P_{\mathcal{X}})$. It is worth noting that we do not assume that $f_{\mathcal{C}} = f_u$ for some $u \in \operatorname{dom} G$, since the infimum of R on $A(\operatorname{dom} G)$ may not be attained. A *consistent learning scheme* generates a random variable $\widehat{u}_{n,\lambda_n}(Z_n)$, taking values in $\ell^2(\mathbb{K})$, from n i.i.d. observations $Z_n = (X_i, Y_i)_{1\leqslant i\leqslant n}$, so that the resulting sequence of random functions $(\hat{f}_n)_{n\in\mathbb{N}} = (A\widehat{u}_{n,\lambda_n}(Z_n))_{n\in\mathbb{N}}$ is *weakly consistent* in the sense that

$$R(\hat{f}_n) \to \inf R(\mathcal{C}) \ \text{ in probability} \quad \Leftrightarrow \quad \|\hat{f}_n - f_{\mathcal{C}}\|_{L^2} \to 0 \text{ in probability,} \tag{2.10}$$

or *strongly consistent* in the sense that

$$R(\hat{f}_n) \to \inf R(\mathcal{C}) \quad \mathsf{P}\text{-a.s.} \quad \Leftrightarrow \quad \|\hat{f}_n - f_{\mathcal{C}}\|_{L^2} \to 0 \quad \mathsf{P}\text{-a.s.}, \tag{2.11}$$

depending on the assumption on the regularization parameters $(\lambda_n)_{n\in\mathbb{N}}$.

Next, we first state our consistency result and then present an algorithm to compute the proposed estimators.

Theorem 2.4 *Suppose that Assumption 2.1 is in force and let $f_{\mathcal{C}}$ be defined as in Proposition 2.3. Let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence in $]0, +\infty[$ converging to 0 and, for every $n \in \mathbb{N}$, let $\hat{f}_n = A\widehat{u}_{n,\lambda_n}(Z_n)$. Then the following hold:*

(i) *Suppose that $\varepsilon(\lambda_n)/\lambda_n^{4/r} \to 0$ and that $1/(\lambda_n^{2/r} n^{1/2}) \to 0$. Then $(\hat{f}_n)_{n\in\mathbb{N}}$ is weakly consistent, i.e., $\|\hat{f}_n - f_{\mathcal{C}}\|_{L^2} \to 0$ in probability.*

(ii) *Suppose that $\varepsilon(\lambda_n) = O(1/n)$ and that $(\log n)/(\lambda_n^{2/r} n^{1/2}) \to 0$. Then $(\hat{f}_n)_{n\in\mathbb{N}}$ is strongly consistent, i.e., $\|\hat{f}_n - f_{\mathcal{C}}\|_{L^2} \to 0$ P-a.s.*

(iii) *Suppose that $f_{\mathcal{C}} \in A(\operatorname{dom} G)$ and set $S = \operatorname{Argmin}_{\operatorname{dom} G} F$. Then there exists a unique $u^\dagger \in S$ which minimizes G over S and $Au^\dagger = f_{\mathcal{C}}$. Moreover, the following hold:*

   (a) *Suppose that $\varepsilon(\lambda_n)/\lambda_n^2 \to 0$ and that $1/(\lambda_n n^{1/2}) \to 0$. Then*

$$\|\widehat{u}_{n,\lambda_n}(Z_n) - u^\dagger\|_r \to 0 \quad \textit{in probability}. \tag{2.12}$$

(b) *Suppose that* $\varepsilon(\lambda_n) = O(1/n)$ *and that* $(\log n)/(\lambda_n n^{1/2}) \to 0$. *Then*

$$\|\widehat{u}_{n,\lambda_n}(Z_n) - u^\dagger\|_r \to 0 \quad \mathsf{P}\text{-}a.s. \tag{2.13}$$

Remark 2.5

(i) In Theorem 2.4(i)-(ii) the weakest conditions on the regularization parameters $(\lambda_n)_{n\in\mathbb{N}}$ occur when $r = 2$. In the case considered in (iii), the consistency conditions do not depend on the exponent r.

(ii) In the special case when, in (1.3), for every $k \in \mathbb{K}$, $h_k = \eta|\cdot|^r$, $C_k = \mathbb{R}$, $D_k = [-\omega_k, \omega_k]$, for some $\omega_k \in \mathbb{R}_{++}$, we recover the elastic net framework of [16] and the same consistency conditions as in [16, Theorem 2 and Theorem 3]. This special case yields a strongly convex problem. In our general setting, the exponent r may take any value in $]1,2]$ and the objective function is only totally convex on bounded sets (see Lemma 3.1). Note also that our framework allows for the enforcement of hard constraints.

(iii) Under the hypotheses of (iii), the consistency extends to the sequence of coefficients $(\widehat{u}_{n,\lambda_n}(Z_n))_{n\in\mathbb{N}}$. This is relevant when one requires the estimators to mimick the properties of $u^\dagger$.

(iv) When $\mathbb{K}$ is finite and, for every $k \in \mathbb{K}$, $g_k = |\cdot|^r$, [23] provides an excess risk bound depending on the cardinality of $\mathbb{K}$ and the level of sparsity of $u^\dagger$ (see also [20]). The case $r = 1$ has been considered in [14]. Appendix B collects useful properties of the proximity operators of power functions.

We now address the algorithmic aspects. The objective function in (2.6) consists of a smooth (quadratic) data fitting term and a separable nondifferentiable term, penalizing each dictionary coefficient individually. Thus a natural choice is to consider the forward-backward splitting algorithm [12]. We stress that, since ε-minimizers are employed in (2.6), algorithms that provide minimizing sequences are necessary. However, when convergence in objective function values is in order, the current theory is not completely satisfying. Indeed, the available results consider only the error free-case and the unrelaxed version [4, 15]. In [28], errors are considered, but only ergodic convergence is proved. In the Theorem 2.7 below, we fill this gap by proving an $o(1/m)$ rate of convergence in objective values with relaxation and in the presence of the following type of errors.

Definition 2.6 Let $\mathcal{H}$ be a real Hilbert space, let $\varphi \in \Gamma_0(\mathcal{H})$, let $(u, w) \in \mathcal{H}^2$, and let $\delta \in \mathbb{R}_+$. The notation $u \simeq_\delta \operatorname{prox}_\varphi w$ means that

$$\varphi(u) + \frac{1}{2}\|u - w\|^2_{\mathcal{H}} \leqslant \min_{v\in\mathcal{H}} \Big(\varphi(v) + \frac{1}{2}\|v - w\|^2_{\mathcal{H}}\Big) + \frac{\delta^2}{2}. \tag{2.14}$$

Theorem 2.7 *Let $\mathcal{H}$ be a real Hilbert space, let $F\colon \mathcal{H} \to \mathbb{R}$ be a convex function which is differentiable on $\mathcal{H}$ with a β-Lipschitz continuous gradient for some $\beta \in \mathbb{R}_{++}$. Let $G \in \Gamma_0(\mathcal{H})$, set $J = F + G$, and suppose that* $\operatorname{Argmin} J \neq \varnothing$. *Let $(\gamma_m)_{m\in\mathbb{N}}$ be a sequence in $\mathbb{R}_{++}$ such that $0 < \inf_{m\in\mathbb{N}} \gamma_m \leqslant \sup_{m\in\mathbb{N}} \gamma_m < 2/\beta$, let $(\tau_m)_{m\in\mathbb{N}}$ be a sequence in $]0,1]$, such that $\inf_{m\in\mathbb{N}} \tau_m > 0$. Let $(\delta_m)_{m\in\mathbb{N}}$ be a summable sequence in $\mathbb{R}_+$ and let $(b_m)_{m\in\mathbb{N}}$ be a summable sequence in $\mathcal{H}$. Fix $u_0 \in \mathcal{H}$ and set*

$$\begin{array}{l} \text{for } m = 0, 1, \ldots \\ \left\lfloor \begin{array}{l} v_m \simeq_{\delta_m} \operatorname{prox}_{\gamma_m G}\big(u_m - \gamma_m(\nabla F(u_m) + b_m)\big) \\ u_{m+1} = u_m + \tau_m(v_m - u_m). \end{array} \right. \end{array} \tag{2.15}$$

Then the following hold:

(i) $(u_m)_{m\in\mathbb{N}}$ *converges weakly to a point in* $\operatorname{Argmin} J$.

(ii) *For every* $u \in \operatorname{Argmin} J$, $\sum_{m\in\mathbb{N}} \|\nabla F(u_m) - \nabla F(u)\|^2 < +\infty$.

(iii) $\sum_{m\in\mathbb{N}} \|v_m - u_m\|^2 < +\infty$.

(iv) $J(u_m) \to \inf J(\mathcal{H})$ *and* $\sum_{m\in\mathbb{N}} |J(v_m) - \inf J(\mathcal{H})|^2 < +\infty$.

(v) *Suppose that* $\sum_{m\in\mathbb{N}}(1-\tau_m) < +\infty$. *Then*

$$\sum_{m\in\mathbb{N}} \big(J(v_m) - \inf J(\mathcal{H})\big) < +\infty \quad \textit{and} \quad \sum_{m\in\mathbb{N}} \big(J(u_m) - \inf J(\mathcal{H})\big) < +\infty.$$

(vi) *Suppose that* $\sum_{m\in\mathbb{N}}(1-\tau_m) < +\infty$, $\sum_{m\in\mathbb{N}} m\delta_m < +\infty$, *and* $\sum_{m\in\mathbb{N}} m\|b_m\| < +\infty$. *Then* $J(u_m) - \inf J(\mathcal{H}) = o(1/m)$.

Remark 2.8 In [4], the rate $O(1/m)$ for objective values is proved in the error-free case and no relaxations ($\delta_m \equiv 0$ and $\tau_m \equiv 1$), assuming that $F+G$ is coercive. On the other hand, an $o(1/m)$ rate on the objective values was derived in [15] in the special case of fixed proximal parameter $\gamma \in \left]0, 2/\beta\right[$, no relaxation, and no errors.

We now propose the following inexact forward-backward algorithm to solve problem 1.2.

Algorithm 2.9 Let $(\gamma_m)_{m\in\mathbb{N}}$ be a sequence in $\mathbb{R}_{++}$ such that $0 < \inf_{m\in\mathbb{N}} \gamma_m \leqslant \sup_{m\in\mathbb{N}} \gamma_m < \lambda/\kappa^2$, let $(\tau_m)_{m\in\mathbb{N}}$ be a sequence in $\left]0,1\right]$ such that $\inf_{m\in\mathbb{N}} \tau_m > 0$. Let $(b_m)_{m\in\mathbb{N}} = ((\beta_{m,k})_{k\in\mathbb{K}})_{m\in\mathbb{N}} \in (\ell^2(\mathbb{K}))^{\mathbb{N}}$ be such that $\sum_{m\in\mathbb{N}} \|b_m\| < +\infty$, let $\zeta \in \mathbb{R}_{++}$, let $p \in \left]1,+\infty\right[$, and let $(\xi_k)_{k\in\mathbb{K}} \in \ell^1(\mathbb{K})$. Fix $(\mu_{0,k})_{k\in\mathbb{K}} \in \ell^2(\mathbb{K})$ and iterate

$$
\begin{array}{l}
\text{for } m = 0,1,\ldots \\
\left\lfloor
\begin{array}{l}
\text{for every } k \in \mathbb{K} \\
\left\lfloor
\begin{array}{l}
\chi_{m,k} = \mu_{m,k} - \dfrac{\gamma_m}{\lambda}\left(\dfrac{2}{n}\displaystyle\sum_{i=1}^{n}\Big(\sum_{j\in\mathbb{K}} \mu_{m,j}\phi_j(x_i) - y_i\Big)\phi_k(x_i) + \beta_{m,k}\right) \\[2ex]
|\alpha_{m,k}| \leqslant \dfrac{\zeta m^{-2p}\xi_k}{4\gamma_m \max\{h_k(|\chi_{m,k}|+2), h_k(-|\chi_{m,k}|-2)\} + 2|\chi_{m,k}| + 1} \\[2ex]
\pi_{m,k} = \operatorname{prox}_{\gamma_m h_k}\big(\operatorname{soft}_{\gamma_m D_k}\chi_{m,k}\big) + \alpha_{m,k} \\[2ex]
\nu_{m,k} = \operatorname{proj}_{C_k}\big(\operatorname{sign}(\chi_{m,k})\max\{0, \operatorname{sign}(\chi_{m,k})\pi_{m,k}\}\big) \\[2ex]
\mu_{m+1,k} = \mu_{m,k} + \tau_m(\nu_{m,k} - \mu_{m,k}).
\end{array}
\right.
\end{array}
\right.
\end{array}
\tag{2.16}
$$

An attractive feature of Algorithm 2.9 is that, at each iteration, each component of the functions in (2.5) is activated componentwise and individually.

Remark 2.10 Nesterov-like [25] variants of the forward-backward splitting algorithm may also be suitable for computing the estimators (2.6) to the extent that they also generate minimizing sequences [28, 30].

Theorem 2.11 *Suppose that Assumption 2.1 is in force. Call*

$$J\colon \ell^2(\mathbb{K}) \to \left]-\infty,+\infty\right] : u = (\mu_k)_{k\in\mathbb{K}} \mapsto \frac{1}{n}\sum_{i=1}^{n} |f_u(x_i) - y_i|^2 + \lambda \sum_{k\in\mathbb{K}} g_k(\mu_k) \tag{2.17}$$

the objective function in (2.6), and let $(u_m)_{m\in\mathbb{N}} = ((\mu_{m,k})_{k\in\mathbb{K}})_{m\in\mathbb{N}}$ and $(v_m)_{m\in\mathbb{N}} = ((\nu_{m,k})_{k\in\mathbb{K}})_{m\in\mathbb{N}}$ be the sequences generated by Algorithm 2.9. Then the following hold:

(i) *J has a unique minimizer $\widehat{u}$, and $\widehat{u} \in \ell^r(\mathbb{K})$.*

(ii) *$\sum_{m\in\mathbb{N}} |J(v_m) - \inf J(\mathcal{H})|^2 < +\infty$, $J(u_m) \to \inf J(\mathcal{H})$, $\|v_m - \widehat{u}\|_r \to 0$, and $\|u_m - \widehat{u}\|_r \to 0$ as $m \to +\infty$. Moreover*

$$\|v_m - \widehat{u}\|_r = O\Big(\sqrt{J(v_m) - \inf J(\mathcal{H})}\Big) \tag{2.18}$$

and

$$\|u_m - \widehat{u}\|_r = O\Big(\sqrt{J(u_m) - \inf J(\mathcal{H})}\Big). \tag{2.19}$$

(iii) *Suppose that $\sum_{m\in\mathbb{N}}(1-\tau_m) < +\infty$. Then*

$$\sum_{m\in\mathbb{N}} \big(J(v_m) - \inf J(\mathcal{H})\big) < +\infty \text{ and } \sum_{m\in\mathbb{N}} \big(J(u_m) - \inf J(\mathcal{H})\big) < +\infty.$$

(iv) *Suppose that $p > 2$, that $\sum_{m\in\mathbb{N}}(1-\tau_m) < +\infty$ and $\sum_{m\in\mathbb{N}} m\|b_m\| < +\infty$. Then*

$$J(u_m) - \inf J(\mathcal{H}) = o(1/m) \quad \text{and} \quad \|u_m - \widehat{u}\|_r = o\big(1/\sqrt{m}\big). \tag{2.20}$$

Remark 2.12

(i) In Algorithm 2.9, the computation of $\operatorname{prox}_{\gamma_m h_k}$ tolerates an error $\alpha_{m,k}$. This is necessary since, in general, the proximity operator is not computable explicitly. In such instances, $\operatorname{prox}_{\gamma_m h_k}$ must be computed iteratively and the bound on $|\alpha_{m,k}|$ in Algorithm 2.9 gives an explicit stopping rule for the iterations.

(ii) The soft-thresholding operator with respect to a bounded interval $D_k = [\underline{\omega}_k, \overline{\omega}_k] \subset \mathbb{R}$ is

$$(\forall \mu \in D_k) \quad \operatorname{soft}_{D_k} \mu = \begin{cases} \mu - \overline{\omega}_k & \text{if } \mu > \overline{\omega}_k \\ 0 & \text{if } \mu \in D_k \\ \mu - \underline{\omega}_k & \text{if } \mu < \underline{\omega}_k. \end{cases} \tag{2.21}$$

The freedom in the choice of the intervals $(D_k)_{k\in\mathbb{K}}$, $(C_k)_{k\in\mathbb{K}}$, and of the exponent r provides flexibility in setting the type of thresholding operation. It is in particular possible to promote selective sparsity. For instance, taking $0 = \underline{\omega}_k < \overline{\omega}_k$ only the positive coefficients are thresholded. Figures 1 and 2 show a few examples.

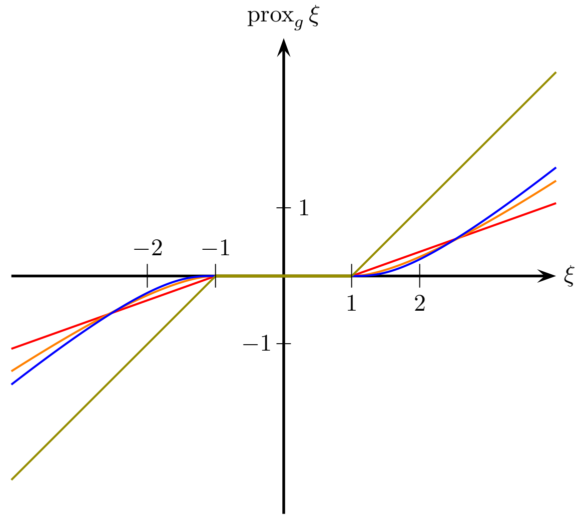

Figure 1: Soft thresholding (green) and prox_g for $g = |\cdot| + 0.9|\cdot|^r$, with $r = 2$ (red), $r = 3/2$ (orange), $r = 4/3$ (blue).

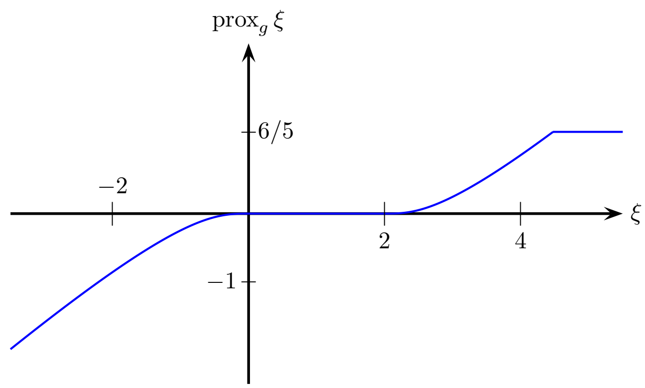

Figure 2: prox_g for $g = \iota_{]-\infty,6/5]} + \sigma_{[0,2]} + 0.9|\cdot|^{4/3}$.

3 Statistical modeling and analysis

Throughout the section Assumption 2.1 is made. Our main objective is to prove Theorem 2.4.

The following result establishes that G is totally convex on bounded sets in $\ell^r(\mathbb{K})$ and gives an explicit lower bound for the relative modulus of total convexity.

Lemma 3.1 *Suppose that Assumption 2.1 is in force. Let $\rho \in \mathbb{R}_{++}$, let $u_0 \in \ell^r(\mathbb{K})$ be such that $\|u_0\|_r \leqslant \rho$, let $u_0^* \in \partial G(u_0)$, and set $M = (7/32)r(r-1)(1-(2/3)^{r-1})$. Then*

$$(\forall\, u \in \ell^r(\mathbb{K})) \quad G(u) - G(u_0) \geqslant \langle u - u_0 \mid u_0^* \rangle + \frac{\eta M \|u - u_0\|_r^2}{(\rho + \|u - u_0\|_r)^{2-r}}. \tag{3.1}$$

Proof. Let $G_|$ be the restriction of G to $\ell^r(\mathbb{K})$, endowed with the norm $\|\cdot\|_r$. Since $u_0 \in \ell^r(\mathbb{K})$

and $u_0^* \in \ell^{r^*}(\mathbb{K})$, we have that $u_0^* \in \partial G_|(u_0)$. Let ψ be the modulus of total convexity of $G_|$ and let φ be the modulus of total convexity of $\|\cdot\|_r^r$ in $\ell^r(\mathbb{K})$. Then, for every $u \in \ell^r(\mathbb{K})$, $G(u) - G(u_0) \geqslant \langle u - u_0, u_0^*\rangle + \psi(u_0; \|u - u_0\|_r)$. Moreover, since $G_| = H + \eta\|\cdot\|_r^r$, with $H \in \Gamma_0(\ell^r(\mathbb{K}))$ (see Lemma A.1), we have $\psi \geqslant \eta\varphi$. The statement follows from [11, Proposition A.9-Remark A.10]. □

The next proposition revisits some results of [2] about Tikhonov-like regularization specialized to our setting.

Proposition 3.2 *Suppose that Assumption 2.1 is in force. For every $(\lambda, \epsilon) \in \mathbb{R}^2_{++}$, let $u_{\lambda,\epsilon}$ be an ε-minimizer of $F + \lambda G$ and let u_G be the minimizer of G. Then the following hold:*

(i) $\inf R(\mathcal{C}) = \inf F(\operatorname{dom} G)$.

(ii) $(\forall (\lambda, \epsilon) \in \mathbb{R}^2_{++})$, $\|u_{\lambda,\epsilon} - u_G\|_r \leqslant \max\big\{\|u_G\|_r, \big(2(F(u_G) + \epsilon)/(\eta M\lambda)\big)^{1/r}\big\}$.

(iii) $F(u_{\lambda,\epsilon}) \to \inf F(\operatorname{dom} G)$ *as* $(\lambda, \epsilon) \to (0^+, 0^+)$.

(iv) *Suppose that* $S = \operatorname{Argmin}_{\operatorname{dom} G} F \neq \varnothing$ *and let* $\varepsilon\colon \mathbb{R}_{++} \to [0, 1]$ *be such that* $\varepsilon(\lambda)/\lambda \to 0^+$ *as* $\lambda \to 0$*. Then there exists* $u^\dagger \in \ell^r(\mathbb{K})$ *such that* $S = \{u^\dagger\}$ *and* $u_{\lambda,\varepsilon(\lambda)} \to u^\dagger$ *as* $\lambda \to 0^+$.

Proof. We first note that it follows from Remark 2.2(ii) that G has a minimizer.

(i): Let $u = (\mu_k)_{k\in\mathbb{K}} \in \ell^2(\mathbb{K}) \cap \bigtimes_{k\in\mathbb{K}} C_k$ and take $\delta \in \mathbb{R}_{++}$. Then there exists a finite set $\mathbb{K}_1 \subset \mathbb{K}$ such that $\sum_{k\in\mathbb{K}\smallsetminus\mathbb{K}_1} |\mu_k|^2 < \delta^2$. Now let $v = (\nu_k)_{k\in\mathbb{K}}$ be such that, for every $k \in \mathbb{K}_1$, $\nu_k = \mu_k$ and, for every $k \in \mathbb{K}\smallsetminus\mathbb{K}_1$, $\nu_k = 0$. We have $v \in \operatorname{dom} G$ and $\|Au - Av\|_{L^2} < \|A\|\delta$. Thus, $\mathcal{C} = \overline{A(\operatorname{dom} G)}$ and the statement follows.

(ii): Let $(\lambda, \varepsilon) \in \mathbb{R}^2_{++}$. We derive from the definition of $u_{\lambda,\varepsilon}$, that $F(u_{\lambda,\varepsilon}) + \lambda G(u_{\lambda,\varepsilon}) \leqslant F(u_G) + \lambda G(u_G) + \varepsilon$ hence, since $0 \in \partial G(u_G)$, it follows from Lemma 3.1 that

$$\frac{\eta M\|u_{\lambda,\epsilon} - u_G\|_r^2}{\big(\|u_G\|_r + \|u_{\lambda,\epsilon} - u_G\|_r\big)^{2-r}} \leqslant G(u_{\lambda,\epsilon}) - G(u_G) \leqslant \frac{F(u_G) + \epsilon}{\lambda}. \tag{3.2}$$

If $\|u_{\lambda,\epsilon} - u_G\|_r \geqslant \|u_G\|_r$, then

$$\frac{\eta M\|u_{\lambda,\epsilon} - u_G\|_r^2}{\big(\|u_G\|_r + \|u_{\lambda,\epsilon} - u_G\|_r\big)^{2-r}} \geqslant \frac{\eta M\|u_{\lambda,\epsilon} - u_G\|_r^2}{\big(2\|u_{\lambda,\epsilon} - u_G\|_r\big)^{2-r}} \geqslant \frac{\eta M}{2}\|u_{\lambda,\epsilon} - u_G\|_r^r \tag{3.3}$$

and hence $\|u_{\lambda,\epsilon} - u_G\|_r^r \leqslant 2(F(u_G) + \epsilon)/(\eta M\lambda)$.

(iii): Let $u \in \operatorname{dom} G$. Then, for every $(\lambda, \epsilon) \in \mathbb{R}^2_{++}$,

$$\begin{aligned}\inf F(\operatorname{dom} G) &\leqslant F(u_{\lambda,\epsilon})\\ &\leqslant F(u_{\lambda,\epsilon}) + \lambda\big(G(u_{\lambda,\epsilon}) - G(u_G)\big)\\ &\leqslant F(u) + \lambda\big(G(u) - G(u_G)\big) + \epsilon.\end{aligned} \tag{3.4}$$

Hence

$$
\begin{aligned}
\inf F(\operatorname{dom} G) &\leqslant \varliminf_{(\lambda,\epsilon)\to(0,0)} F(u_{\lambda,\epsilon}) \\
&\leqslant \varlimsup_{(\lambda,\epsilon)\to(0,0)} F(u_{\lambda,\epsilon}) \\
&\leqslant \varlimsup_{(\lambda,\epsilon)\to(0,0)} \big(F(u) + \lambda\big(G(u) - G(u_G)\big) + \epsilon\big) \\
&\leqslant F(u).
\end{aligned}
\tag{3.5}
$$

Since u is arbitrary, the statement follows.

(iv): Since S is convex and $G \in \Gamma_0(\ell^2(\mathbb{K}))$ is strictly convex, coercive, and $\operatorname{dom} G \subset \ell^r(\mathbb{K})$, it follows from [3, Corollary 11.15(ii)] that there exists $u^\dagger \in \ell^r(\mathbb{K})$ such that $S = \{u^\dagger\}$. Moreover,

$$
G(u_{\lambda,\epsilon(\lambda)}) \leqslant (F(u^\dagger) - F(u_{\lambda,\epsilon(\lambda)}) + \epsilon(\lambda))/\lambda + G(u^\dagger) \leqslant G(u^\dagger) + \epsilon(\lambda)/\lambda, \tag{3.6}
$$

which implies that $(G(u_{\lambda,\epsilon(\lambda)}))_{\lambda\in\mathbb{R}_{++}}$ is bounded. Since G is coercive, the family $(u_{\lambda,\varepsilon(\lambda)})_{\lambda\in\mathbb{R}_{++}}$ is bounded as well. We deduce from [33, Proposition 3.6.5] (see also [6]) that there exists an increasing function $\phi\colon \mathbb{R}_{++} \to \mathbb{R}_+$ such that $\phi(0) = 0$, for every $t \in \mathbb{R}_{++}$, $\phi(t) > 0$, and

$$
(\forall \lambda \in \mathbb{R}_{++}) \quad \phi\Big(\frac{\|u_{\lambda,\varepsilon(\lambda)} - u^\dagger\|}{2}\Big) \leqslant \frac{G(u^\dagger) + G(u_{\lambda,\varepsilon(\lambda)})}{2} - G\Big(\frac{u_{\lambda,\varepsilon(\lambda)} + u^\dagger}{2}\Big). \tag{3.7}
$$

Hence, arguing as in [8, Proof of Proposition 3.1(vi)], we obtain $u_\lambda \to u^\dagger$ as $\lambda \to 0^+$. □

Next, we give a representer and stability theorem which generalizes existing results [17, 29] to our class of regularization functions.

Theorem 3.3 *Suppose that Assumption 2.1 is in force. Set $M = (7/32)r(r-1)(1-(2/3)^{r-1})$, let $\lambda \in \mathbb{R}_{++}$, and let $u_\lambda \in \ell^r(\mathbb{K})$ be the minimizer of $F + \lambda G$. Then the following hold:*

(i) *The function*

$$
\Psi_\lambda\colon \mathcal{X} \times \mathcal{Y} \to \ell^2(\mathbb{K})\colon (x, y) \mapsto 2(f_{u_\lambda}(x) - y)\Phi(x) \tag{3.8}
$$

is bounded and $\|\Psi_\lambda\|_\infty \leqslant 2\kappa(\kappa\|u_\lambda\|_2 + b)$. Moreover $\|\Psi_\lambda\|_2 \leqslant 2\kappa\sqrt{R(f_{u_\lambda})}$ and $-\mathsf{E}_P(\Psi_\lambda) \in \lambda\partial G(u_\lambda)$.

(ii) *Let $n \in \mathbb{N}^*$. Then there exists $\widehat{v} \in \ell^r(\mathbb{K})$ such that $\|\widehat{v} - \widehat{u}_{n,\lambda}(z_n)\|_r \leqslant \sqrt{\varepsilon(\lambda)}$*

$$
\frac{\eta M\|\widehat{v} - u_\lambda\|_r}{(\|u_\lambda\|_r + \|\widehat{v} - u_\lambda\|_r)^{2-r}} \leqslant \frac{1}{\lambda}\Big(\Big\|\frac{1}{n}\sum_{i=1}^n \Psi_\lambda(x_i, y_i) - \mathsf{E}_P(\Psi_\lambda)\Big\|_2 + \sqrt{\varepsilon(\lambda)}\Big). \tag{3.9}
$$

Proof. (i): First note that [3, Corollary 11.15(ii)] asserts that u_λ is well defined, since $F + \lambda G$ is proper and lower semicontinuous, and, by Remark 2.2(ii), strictly convex and coercive. Furthermore [3, Corollary 26.3(vii)] implies that $-\nabla F(u_\lambda) \in \lambda\partial G(u_\lambda)$. We derive from (2.3) that $A^*\colon L^2(P_{\mathcal{X}}) \to \ell^2(\mathbb{K})\colon f \mapsto \mathsf{E}_{P_{\mathcal{X}}}(f\Phi)$, and hence, since $F = R \circ A$,

$$
(\forall u \in \ell^2(\mathbb{K})) \quad \nabla F(u) = A^*\nabla R(f_u) = E_P(\varphi), \tag{3.10}
$$

where $\varphi\colon (x,y) \mapsto 2\big(f_u(x) - y)\Phi(x)\big)$. Let $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Then

$$|f_{u_\lambda}(x) - y| \leqslant |f_{u_\lambda}(x)| + |y| \leqslant \sum_{k\in\mathbb{K}} |\langle u_\lambda \mid e_k\rangle||\phi_k(x)| + b \leqslant \kappa\|u_\lambda\|_2 + b \tag{3.11}$$

and hence $\|\Psi_\lambda(x,y)\|_2 \leqslant 2|f_{u_\lambda}(x) - y|\|\Phi(x)\|_2 \leqslant 2(\kappa\|u_\lambda\|_2 + b)\kappa$. Moreover,

$$\int_{\mathcal{X}\times\mathcal{Y}} \|\Psi_\lambda(x,y)\|_2^2 \mathrm{d}P(x,y) \leqslant \int_{\mathcal{X}\times\mathcal{Y}} \big(2\kappa|f_{u_\lambda}(x) - y|\big)^2 \mathrm{d}P(x,y) = 4\kappa^2 R(f_{u_\lambda}). \tag{3.12}$$

(ii): Let $\widehat{F}_n\colon \ell^2(\mathbb{K}) \to \mathbb{R}_+\colon u \mapsto (1/n)\sum_{i=1}^n |f_u(x_i) - y_i|^2$. Since the restriction of G to $\ell^r(\mathbb{K})$ is in $\Gamma_0(\ell^r(\mathbb{K}))$ by Lemma A.1, Ekeland's variational principle [3, Theorem 1.45] implies that there exists $\widehat{v} \in \ell^r(\mathbb{K})$ such that $\|\widehat{u}_{n,\lambda}(z_n) - \widehat{v}\|_r \leqslant \sqrt{\varepsilon(\lambda)}$ and $\inf \|\partial(\widehat{F}_n + \lambda G)(\widehat{v})\|_{r^*} \leqslant \sqrt{\varepsilon(\lambda)}$. Using the inequality $a^2 - b^2 \geqslant 2(a-b)b$, we derive from definitions (3.8) and (2.4) that, for every $i \in \{1,\ldots,n\}$,

$$\begin{aligned}\sum_{k\in\mathbb{K}} \langle \widehat{v} - u_\lambda \mid e_k\rangle\langle\Psi_\lambda(x_i,y_i) \mid e_k\rangle &= \sum_{k\in\mathbb{K}} \langle \widehat{v} - u_\lambda \mid e_k\rangle 2(f_{u_\lambda}(x_i) - y_i)\phi_k(x_i)\\ &= 2(f_{\widehat{v}}(x_i) - f_{u_\lambda}(x_i))(f_{u_\lambda}(x_i) - y_i)\\ &\leqslant (y_i - f_{\widehat{v}}(x_i))^2 - (y_i - f_{u_\lambda}(x_i))^2 \end{aligned} \tag{3.13}$$

and, summing over i and dividing by n, we obtain

$$\widehat{F}_n(\widehat{v}) - \widehat{F}_n(u_\lambda) \geqslant \sum_{k\in\mathbb{K}} \langle \widehat{v} - u_\lambda \mid e_k\rangle \Big\langle \frac{1}{n}\sum\nolimits_{i=1}^n \Psi_\lambda(x_i,y_i) \,\Big|\, e_k\Big\rangle. \tag{3.14}$$

Lemma 3.1 and (i) yield

$$\lambda G(\widehat{v}) - \lambda G(u_\lambda) \geqslant \langle \widehat{v} - u_\lambda \mid -\mathsf{E}_P(\Psi_\lambda)\rangle + \lambda\eta M \frac{\|\widehat{v} - u_\lambda\|_r^2}{(\|u_\lambda\|_r + \|\widehat{v} - u_\lambda\|_r)^{2-r}}. \tag{3.15}$$

Next, since $\inf \|\partial(\widehat{F}_n + \lambda G)(\widehat{v})\|_{r^*} \leqslant \sqrt{\varepsilon(\lambda)}$, there exists $\widehat{e}^* \in \ell^{r^*}(\mathbb{K})$ such that $\|\widehat{e}^*\|_{r^*} \leqslant \sqrt{\varepsilon(\lambda)}$ and $\langle u_\lambda - \widehat{v} \mid \widehat{e}^*\rangle \leqslant (\widehat{F}_n + \lambda G)(u_\lambda) - (\widehat{F}_n + \lambda G)(\widehat{v})$. Summing inequalities (3.14) and (3.15), we have

$$\begin{aligned}\sqrt{\varepsilon(\lambda)}\|\widehat{v} - u_\lambda\|_r &\geqslant (\widehat{F}_n + \lambda G)(\widehat{v}) - (\widehat{F}_n + \lambda G)(u_\lambda)\\ &\geqslant \sum_{k\in\mathbb{K}} \langle \widehat{v} - u_\lambda \mid e_k\rangle \Big\langle \frac{1}{n}\sum_{i=1}^n \Psi_\lambda(x_i,y_i) - \mathsf{E}_P(\Psi_\lambda) \,\Big|\, e_k\Big\rangle\\ &\quad + \frac{\lambda\eta M\|\widehat{v} - u_\lambda\|_r^2}{(\|u_\lambda\|_r + \|\widehat{v} - u_\lambda\|_r)^{2-r}}. \end{aligned} \tag{3.16}$$

Hence, using Hölder's inequality,

$$\frac{\lambda\eta M\|\widehat{v} - u_\lambda\|_r^2}{(\|u_\lambda\|_r + \|\widehat{v} - u_\lambda\|_r)^{2-r}} \leqslant \|\widehat{v} - u_\lambda\|_r \bigg(\Big\|\frac{1}{n}\sum\nolimits_{i=1}^n \Psi_\lambda(x_i,y_i) - \mathsf{E}_P(\Psi_\lambda)\Big\|_{r^*} + \sqrt{\varepsilon(\lambda)}\bigg) \tag{3.17}$$

and the statement follows from the fact that $\|\cdot\|_{r^*} \leqslant \|\cdot\|_2$. □

We recall the following concentration inequality in Hilbert spaces [32] and give the proof of the main result of this section.

Lemma 3.4 (Bernstein's inequality) *Let $(U_i)_{1\leqslant i\leqslant n}$ be a finite sequence of i.i.d. random variables on a probability space $(\Omega, \mathfrak{A}, \mathsf{P})$ and taking values in a real separable Hilbert space $\mathcal{H}$. Let $\beta > 0$, let $\sigma > 0$ and suppose that $\max_{1\leqslant i\leqslant n} \|U_i\| \leqslant \beta$ and that $\mathsf{E}_{\mathsf{P}}\|U_i\|^2 \leqslant \sigma^2$. Then for every $\tau > 0$ and every integer $n \geqslant 1$*

$$P\left[\left\|\frac{1}{n}\sum_{i=1}^{n}(U_i - \mathsf{E}_P U_i)\right\| \geqslant \frac{2\sigma}{\sqrt{n}} + 4\sigma\sqrt{\frac{\tau}{n}} + \frac{4\beta\tau}{3n}\right] \leqslant e^{-\tau}. \tag{3.18}$$

Proof. [of Proposition 2.3] (i): For every $f \in \mathcal{C}$, $R(f) = \|f - f^\dagger\|_{L^2}^2 + \inf R(L^2(P_{\mathcal{X}}))$. Therefore, minimizing R over $\mathcal{C}$ turns to find the element of $\mathcal{C}$ which is nearest to $f^\dagger$ in $L^2(P_{\mathcal{X}})$.

(ii): It follows from (i), that $\inf R(\mathcal{C}) = \|f_{\mathcal{C}} - f^\dagger\|_{L^2}^2 + \inf R(L^2(P_{\mathcal{X}}))$. Therefore, since for every $f \in \mathcal{C}$, $\langle f - f_{\mathcal{C}} \mid f^\dagger - f_{\mathcal{C}}\rangle \leqslant 0$, we have $R(f) - \inf R(\mathcal{C}) = \|f - f^\dagger\|_{L^2}^2 - \|f_{\mathcal{C}} - f^\dagger\|_{L^2}^2 = \|f - f_{\mathcal{C}}\|_{L^2}^2 + 2\langle f - f_{\mathcal{C}} \mid f_{\mathcal{C}} - f^\dagger\rangle \geqslant \|f - f_{\mathcal{C}}\|_{L^2}^2$.

(iii): Let $f \in \mathcal{C}$. Using the fact that, for every $(a, b, c) \in \mathbb{R}_+^3$ with $a \leqslant b$, $\sqrt{a+c} - \sqrt{b+c} \leqslant \sqrt{a} - \sqrt{b}$, we derive that

$$\sqrt{R(f)} - \sqrt{\inf R(\mathcal{C})} \leqslant \|f - f^\dagger\|_{L^2} - \|f_{\mathcal{C}} - f^\dagger\|_{L^2} \leqslant \|f - f_{\mathcal{C}}\|_{L^2}. \tag{3.19}$$

Therefore, using the inequality $a^2 - b^2 \leqslant 2a(a-b)$, we obtain

$$\begin{aligned}
R(f) - \inf R(\mathcal{C}) &\leqslant 2\sqrt{R(f)}\,\|f - f_{\mathcal{C}}\|_{L^2}\\
&= 2\sqrt{\left(\|f - f^\dagger\|_{L^2}^2 + \inf R(L^2(\mathcal{X}))\right)}\,\|f - f_{\mathcal{C}}\|_{L^2}\\
&\leqslant 2\Big(\left(\|f - f_{\mathcal{C}}\|_{L^2} + \|f_{\mathcal{C}} - f^\dagger\|_{L^2}\right)^2 + \inf R(L^2(\mathcal{X}))\Big)^{1/2}\|f - f_{\mathcal{C}}\|_{L^2}\\
&= 2\Big(\Big(\|f - f_{\mathcal{C}}\|_{L^2} + \sqrt{\inf_{\mathcal{C}} R - \inf_{L^2(P_{\mathcal{X}})} R}\Big)^2 + \inf R(L^2(\mathcal{X}))\Big)^{1/2}\|f - f_{\mathcal{C}}\|_{L^2}
\end{aligned} \tag{3.20}$$

□

Proof. [of Theorem 2.4] (i): Let $n \in \mathbb{N}^*$, let $z_n = (x_i, y_i)_{1\leqslant i\leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$ and let $\widehat{F}_n \colon u \in \ell^2(\mathbb{K}) \to \mathbb{R}_+ \colon u \mapsto (1/n)\sum_{i=1}^n |f_u(x_i) - y_i|^2$. Let $u_G \in \operatorname{Argmin} G$, let $\lambda \in \mathbb{R}_{++}$, and let $\rho_\lambda = \max\left\{1, \|u_G\|_r, (2(b + \kappa\|u_G\| + 1)^2/(\eta M\lambda))^{1/r}\right\}$. Since $F(u_G) \leqslant (b + \kappa\|u_G\|)^2$ and $\widehat{F}_n(u_G) \leqslant (b + \kappa\|u_G\|)^2$, from the definition of ρ_λ and Proposition 3.2(ii) we derive that $\|u_\lambda - u_G\|_r \leqslant \rho_\lambda$ and $\|\widehat{u}_{n,\lambda}(z_n) - u_G\|_r \leqslant \rho_\lambda$. It follows from Theorem 3.3 that there exist $\Psi_\lambda \colon \mathcal{X} \times \mathcal{Y} \to \ell^2(\mathbb{K})$ and $\widehat{v} \in \ell^r(\mathbb{K})$ such that $\|\widehat{v} - \widehat{u}_{n,\lambda}(z_n)\| \leqslant \sqrt{\varepsilon(\lambda)}$ and

$$\frac{M\eta\|\widehat{v} - u_\lambda\|_r}{(\|u_\lambda\|_r + \|\widehat{v} - u_\lambda\|_r)^{2-r}} \leqslant \frac{1}{\lambda}\left(\left\|\frac{1}{n}\sum_{i=1}^n \Psi_\lambda(x_i, y_i) - \mathsf{E}_P(\Psi_\lambda)\right\|_2 + \sqrt{\varepsilon(\lambda)}\right). \tag{3.21}$$

Therefore

$$\|\widehat{v} - u_\lambda\|_r \leqslant \frac{(4\rho_\lambda)^{2-r}}{M\eta\lambda}\left(\left\|\mathsf{E}_P(\Psi_\lambda) - \frac{1}{n}\sum_{i=1}^n \Psi_\lambda(x_i, y_i)\right\|_2 + \sqrt{\varepsilon(\lambda)}\right). \tag{3.22}$$

Thus,

$$\|\widehat{u}_{n,\lambda}(z_n) - u_\lambda\|_r \leqslant \sqrt{\varepsilon(\lambda)} + \frac{(4\rho_\lambda)^{2-r}}{M\eta\lambda}\left(\left\|\mathsf{E}_P(\Psi_\lambda) - \frac{1}{n}\sum_{i=1}^n \Psi_\lambda(x_i, y_i)\right\|_2 + \sqrt{\varepsilon(\lambda)}\right). \tag{3.23}$$

Now, consider the i.i.d. random vectors $\Psi_\lambda(X_i, Y_i)\colon \Omega \to \ell^2(\mathbb{K})$, for $1 \leqslant i \leqslant n$. It follows from Theorem 3.3(i) that $\max_{1\leqslant i\leqslant n} \|\Psi_\lambda(X_i, Y_i)\| \leqslant 2\kappa(\kappa\rho_\lambda + b)$ and that $\max_{1\leqslant i\leqslant n} \mathsf{E}_\mathsf{P}\|\Psi_\lambda(X_i, Y_i)\|^2 \leqslant 4\kappa^2 R(f_{u_\lambda})$. Now set $\beta_\lambda = 2\kappa(\kappa\rho_\lambda + b)$ and $\sigma_\lambda^2 = \kappa^2 R(f_{u_\lambda})$. Then Bernstein's inequality in Hilbert spaces (Lemma 3.4) gives

$$(\forall \tau \in \mathbb{R}_{++}) \quad \mathsf{P}\Big[\Big\|\mathsf{E}(\Psi_\lambda(X,Y)) - \frac{1}{n}\sum_{i=1}^n \Psi_\lambda(X_i, Y_i)\Big\|_2 \leqslant \delta(n,\lambda,\tau)\Big] \geqslant 1 - e^{-\tau}, \tag{3.24}$$

where $\delta(n,\lambda,\tau) = 2\sigma_\lambda/\sqrt{n} + 4\sigma_\lambda\sqrt{\tau/n} + 4\beta_\lambda\tau/(3n)$. Thus, recalling (3.23) we have

$$\mathsf{P}\Big[\|\widehat{u}_{n,\lambda}(Z_n) - u_\lambda\|_r > \sqrt{\varepsilon(\lambda)} + \frac{(4\rho_\lambda)^{2-r}}{M\eta\lambda}\big(\delta(n,\lambda,\tau) + \sqrt{\varepsilon(\lambda)}\big)\Big] \leqslant e^{-\tau}. \tag{3.25}$$

Set $\gamma_0 = 2(b + \kappa\|u_G\| + 1)^2$ and $\gamma_1 = 4^{2-r}\gamma_0^{2/r-1}/(\eta M)^{2/r}$. We note that, since σ_λ is bounded, say by γ_2, for $\lambda < 1$ sufficiently small, we have

$$\begin{aligned}
&\frac{(4\rho_\lambda)^{2-r}}{M\eta\lambda}\big(\delta(n,\lambda,\tau) + \sqrt{\varepsilon(\lambda)}\big)\\
&= \frac{4^{2-r}}{M\eta}\Big(\frac{\gamma_0}{\eta M}\Big)^{\frac{2}{r}-1}\frac{1}{\lambda^{2/r}}\Big(\frac{2\sigma_\lambda}{\sqrt{n}} + 4\sigma_\lambda\sqrt{\frac{\tau}{n}} + \frac{4\beta_\lambda\tau}{3n} + \sqrt{\varepsilon(\lambda)}\Big)\\
&\leqslant \gamma_1\Big(\frac{2\gamma_2}{\lambda^{2/r}n^{1/2}} + 4\gamma_2\frac{\sqrt{\tau}}{\lambda^{2/r}n^{1/2}} + \frac{8\tau\kappa^2\gamma_0/(\eta M)^{1/r}}{3n\lambda^{3/r}} + \frac{8\tau\kappa b}{3n\lambda^{2/r}} + \frac{\sqrt{\varepsilon(\lambda)}}{\lambda^{2/r}}\Big).
\end{aligned} \tag{3.26}$$

Therefore, since $1/(\lambda_n^{2/r}n^{1/2}) \to 0$ and $\sqrt{\varepsilon(\lambda_n)}/\lambda_n^{2/r} \to 0$ it follows that

$$\frac{(4\rho_{\lambda_n})^{2-r}}{\lambda_n}\Big(\delta(n,\lambda_n,\tau) + \sqrt{\varepsilon(\lambda_n)}\Big) \to 0 \tag{3.27}$$

and hence, in view of (3.25), we get $\|\widehat{u}_{n,\lambda_n}(Z_n) - u_{\lambda_n}\|_r \to 0$ in probability. Moreover, using Proposition 2.3(ii),

$$\begin{aligned}
\|\widehat{f}_n - f_\mathcal{C}\|_{L^2} &\leqslant \|A\widehat{u}_{n,\lambda_n}(Z_n) - Au_{\lambda_n}\|_{L^2} + \|Au_{\lambda_n} - f_\mathcal{C}\|_{L^2}\\
&\leqslant \|A\|\|\widehat{u}_{n,\lambda_n}(Z_n) - u_{\lambda_n}\|_r + \sqrt{F(u_{\lambda_n}) - \inf F(\operatorname{dom} G)}.
\end{aligned} \tag{3.28}$$

Since $F(u_{\lambda_n}) - \inf F(\operatorname{dom} G) \to 0$ by Proposition 3.2(iii), and $\|\widehat{u}_{n,\lambda}(Z_n) - u_{\lambda_n}\|_r \to 0$ in probability, we derive that $\|\widehat{f}_n - f_\mathcal{C}\|_{L^2} \to 0$ in probability.

(ii): Let $n \in \mathbb{N}^*$, let $\eta \in \mathbb{R}_{++}$, and set

$$\Omega_{n,\eta} = \Big\{\|\widehat{f}_n - f_\mathcal{C}\|_{L^2} > \|A\|\eta + \sqrt{F(u_{\lambda_n}) - \inf F(\operatorname{dom} G)}\Big\}. \tag{3.29}$$

Since $\varepsilon(\lambda_n) = O(1/n)$, it follows from (3.26) that there exists $\gamma_3 \in \mathbb{R}_{++}$ such that, for every $\tau \in \left]1, +\infty\right[$, and every $n \in \mathbb{N}^*$,

$$\frac{(4\rho_{\lambda_n})^{2-r}}{\eta M\lambda_n}\big(\delta(n,\lambda_n,\tau) + \sqrt{\varepsilon(\lambda_n)}\big) \leqslant \frac{\gamma_3\tau}{\lambda_n^{2/r}n^{1/2}}. \tag{3.30}$$

Let $\xi \in \left]1, +\infty\right[$. There exists $\overline{n} \in \mathbb{N}^*$, such that, for every integer $n \geqslant \overline{n}$,

$$\frac{\gamma_3}{\lambda_n^{2/r}n^{1/2}} \leqslant \frac{\gamma_3\xi\log n}{\lambda_n^{2/r}n^{1/2}} \leqslant \eta. \tag{3.31}$$

Therefore, it follows from (3.25), (3.28), (3.30), and (3.31) that, for n large enough,

$$\mathsf{P}\Omega_{n,\eta} \leqslant \exp\left(-\frac{\eta\lambda_n^{2/r} n^{1/2}}{\gamma_3}\right) \leqslant \exp\left(-\xi \log n\right) = n^{-\xi}. \tag{3.32}$$

Thus, $\sum_{n=\bar{n}}^{+\infty} \mathsf{P}\Omega_{n,\eta} < +\infty$ and we derive from the Borel-Cantelli lemma that $\mathsf{P}\big(\bigcap_{k\geqslant\bar{n}}\bigcup_{n\geqslant k}\Omega_{n,\eta}\big) = 0$. Recalling Proposition 3.2(iii), we conclude that the sequence $\|\widehat{f}_n - f_{\mathcal{C}}\|_{L^2} \to 0$ P-a.s.

(iii): First note that Proposition 3.2(iii) implies that $u^\dagger$ is well defined and that $\rho = \sup_{\lambda\in\mathbb{R}_{++}}\|u_\lambda\| < +\infty$. Now, let $\lambda \in \mathbb{R}_{++}$ and let $n \in \mathbb{N}^*$. Since $\|u_\lambda\| \leqslant \rho$, arguing as in the proof of (i), we obtain

$$(\forall \tau \in \mathbb{R}_{++})\ \mathsf{P}\Big[\|\widehat{u}_{n,\lambda}(Z_n) - u_\lambda\|_r > \sqrt{\varepsilon(\lambda)} + \frac{(4\rho)^{2-r}}{M\lambda}\big(\delta(n,\tau) + \sqrt{\varepsilon(\lambda)}\big)\Big] \leqslant e^{-\tau}, \tag{3.33}$$

where $\sigma = 2\kappa(\kappa\rho + b)$ and $\delta(n,\tau) = 4\sigma/\sqrt{n} + 4\sigma\sqrt{\tau/n} + 4\sigma\tau/(3n)$.

(iii)(a): Since $1/(\lambda_n n^{1/2}) \to 0$, we have $(1/\lambda_n)\delta(n,\tau) \to 0$ and hence in view of (3.33), $\|\widehat{u}_{n,\lambda_n}(Z_n) - u_{\lambda_n}\|_r \to 0$ in probability. Moreover, since $\|u_{n,\lambda_n}(Z_n) - u^\dagger\| \leqslant \|u_{n,\lambda_n}(Z_n) - u_{\lambda_n}\| + \|u_{\lambda_n} - u^\dagger\|$, the statement follows by Proposition 3.2(iv).

(iii)(b): The proof follows the same line as that of (ii). □

4 Algorithm

The goal of this section is to prove Theorem 2.7 and Theorem 2.11. The proof of Theorem 2.7 is based on the following fact.

Lemma 4.1 [30, Lemma 4.1] *Let $\mathcal{H}$ be a real Hilbert space, let $\beta \in \mathbb{R}_{++}$, and let $\delta \in \mathbb{R}_+$. Let $F\colon \mathcal{H} \to \left]-\infty,+\infty\right]$ be a convex differentiable function with β-Lipschitz continuous gradient, and let $G \in \Gamma_0(\mathcal{H})$. Then, for every $(u,v,w) \in \mathcal{H}^3$ and every $v^* \in \partial_\delta G(v)$,*

$$(F+G)(v) \leqslant (F+G)(u) + \langle v - u \mid \nabla F(w) + v^*\rangle + \frac{\beta}{2}\|v-w\|^2 + \delta. \tag{4.1}$$

Proof. [of Theorem 2.7] Let $m \in \mathbb{N}$ and set

$$\tilde{v}_m = \operatorname{prox}_{\gamma_m G}(u_m - \gamma_m(\nabla F(u_m) + b_m)). \tag{4.2}$$

Since

$$v_m \in \operatorname{Argmin}^{\delta_m^2/(2\gamma_m)}_{w\in\mathcal{H}} \left\{G(w) + \frac{1}{2\gamma_m}\|w - (u_m - \gamma_m(\nabla F(u_m) + b_m))\|^2\right\}, \tag{4.3}$$

using the strong convexity of the objective function in (4.3), we get

$$\|v_m - \tilde{v}_m\| \leqslant \delta_m. \tag{4.4}$$

Therefore, setting $a_m = v_m - \tilde{v}_m$, we have

$$u_{m+1} = u_m + \tau_m\big(\operatorname{prox}_{\gamma_m G}(u_m - \gamma_m(\nabla F(u_m) + b_m)) + a_m - u_m\big). \tag{4.5}$$

Hence (2.15) is an instance of the inexact forward-backward algorithm studied in [12] and we can therefore use the results of [12, Theorem 3.4].

(i)–(ii): The statements follow from [12, Theorem 3.4(i)-(ii)].

(iii): We have

$$\begin{aligned}\|u_m - v_m\|^2 &\leqslant 2\|u_m - \tilde{v}_m\|^2 + 2\|a_m\|^2 \\ &\leqslant 4\|u_m - \operatorname{prox}_{\gamma_m G}(u_m - \gamma_m \nabla F(u_m))\|^2 + 4\|b_m\|^2 + 2\|a_m\|^2. \end{aligned} \tag{4.6}$$

Therefore, the statement follows from [12, Theorem 3.4(iii)].

(iv): By (4.3) and [27, Lemma 1], there exist $\delta_{1,m} \in [0, +\infty[$, $\delta_{2,m} \in [0, +\infty[$, and $e_m \in \mathcal{H}$ with $\delta_{1,m}^2 + \delta_{2,m}^2 \leqslant \delta_m^2$ and $\|e_m\| \leqslant \delta_{2,m}$ such that

$$v_m^* = \frac{u_m - v_m}{\gamma_m} - (\nabla F(u_m) + b_m) + \frac{e_m}{\gamma_m} \in \partial_{\delta_{1,m}^2/(2\gamma_m)} G(v_m). \tag{4.7}$$

Now set $J = F + G$. It follows from Lemma 4.1 that, for every $u \in \mathcal{H}$,

$$\begin{aligned} &J(v_m) - J(u) \\ &\quad\leqslant \langle v_m - u \mid \nabla F(u_m) + v_m^* \rangle + \frac{\beta}{2}\|v_m - u_m\|^2 + \frac{\delta_{1,m}^2}{2\gamma_m} \\ &\quad= \frac{1}{\gamma_m}\langle v_m - u \mid u_m - v_m\rangle + \frac{\beta}{2}\|v_m - u_m\|^2 \\ &\qquad\qquad + \frac{1}{\gamma_m}\langle v_m - u \mid e_m - \gamma_m b_m\rangle + \frac{\delta_{1,m}^2}{2\gamma_m} \\ &\quad= \frac{1}{2\gamma_m}\big(\|u_m - u\|^2 - \|v_m - u\|^2\big) + \frac{1}{2}\left(\beta - \frac{1}{\gamma_m}\right)\|v_m - u_m\|^2 \\ &\qquad\qquad + \frac{1}{\gamma_m}\langle v_m - u \mid e_m - \gamma_m b_m\rangle + \frac{\delta_{1,m}^2}{2\gamma_m}. \end{aligned} \tag{4.8}$$

We derive from (i) and (iii) that $(\langle v_m - u \mid u_m - v_m\rangle)_{m\in\mathbb{N}}$ is square summable. Therefore, if we let $u \in \operatorname{Argmin} J$, it follows from (4.8) that $(J(v_m) - \inf J(\mathcal{H}))_{m\in\mathbb{N}}$ is square summable. Now, if we let $u = u_m$ in (4.8) we have

$$\begin{aligned} J(v_m) - J(u_m) &\leqslant \left(\frac{\beta}{2} - \frac{1}{\gamma_m}\right)\|u_m - v_m\|^2 + \frac{1}{\gamma_m}\left(\|u_m - v_m\|\|e_m - \gamma_m b_m\| + \frac{\delta_{1,m}^2}{2}\right) \\ &\leqslant \frac{1}{\gamma_m}\left(\|u_m - v_m\|\|e_m - \gamma_m b_m\| + \frac{\delta_{1,m}^2}{2}\right). \end{aligned} \tag{4.9}$$

Set $\underline{\gamma} = \inf_{m\in\mathbb{N}} \gamma_m$. Since $u_{m+1} = u_m + \tau_m(v_m - u_m)$, using the convexity of J and (4.9), we get

$$\begin{aligned} J(u_{m+1}) - \inf J(\mathcal{H}) &\leqslant J(u_m) - \inf J(\mathcal{H}) + \tau_m(J(v_m) - J(u_m)) \\ &\leqslant J(u_m) - \inf J(\mathcal{H}) + \underline{\gamma}^{-1}\big(\|u_m - v_m\|\|e_m - \gamma_m b_m\| + \delta_{1,m}^2/2\big). \end{aligned} \tag{4.10}$$

Thus, since $\big(\|u_m - v_m\|\,\|e_m - \gamma_m b_m\| + \delta_{1,m}^2/2\big)_{m\in\mathbb{N}}$ is summable, [26, Lemma 2.2.2], ensures that $(J(u_m) - \inf J(\mathcal{H}))_{m\in\mathbb{N}}$ converges. In view of the inequalities in (4.10) that its limit must be 0.

(v): Let $u \in \operatorname{Argmin} J$. Since, $u_m - u = (1-\tau_{m-1})(u_{m-1}-u) + \tau_{m-1}(v_{m-1}-u)$, it follows from the convexity of $\|\cdot\|^2$ that

$$\|u_m - u\| - \|v_m - u\|^2 \leqslant (1-\tau_{m-1})\|u_{m-1}-u\|^2 + \|v_{m-1}-u\|^2 - \|v_m - u\|^2. \tag{4.11}$$

Therefore, it follows from (4.8) that

$$\begin{aligned} 0 &\leqslant J(v_m) - J(u) \\ &\leqslant \frac{1-\lambda_{m-1}}{2\underline{\gamma}}\|u_{m-1}-u\|^2 + \frac{1}{2\underline{\gamma}}\big(\|v_{m-1}-u\|^2 - \|v_m-u\|^2\big) \\ &+ \frac{1}{2}\left(\beta - \frac{1}{\gamma_m}\right)\|v_m - u_m\|^2 + \frac{1}{\underline{\gamma}}\left(\|v_m-u\|\|e_m - \gamma_m b_m\| + \frac{\delta_{1,m}^2}{2}\right). \end{aligned} \tag{4.12}$$

Hence, $(J(v_m) - \inf J(\mathcal{H}))_{m\in\mathbb{N}}$ is summable, for each term on the right hand side of (4.12) is summable. Since $u_{m+1} = (1-\tau_m)u_m + \tau_m v_m$, convexity of J yields

$$0 \leqslant J(u_{m+1}) - \inf J(\mathcal{H}) \leqslant (1-\tau_m)\big(J(u_m) - \inf J(\mathcal{H})\big) + \tau_m\big(J(v_m) - \inf J(\mathcal{H})\big). \tag{4.13}$$

The summability of $(1-\tau_m)_{m\in\mathbb{N}}$ and $\big(J(v_m) - \inf J(\mathcal{H})\big)_{m\in\mathbb{N}}$ implies that of $\big(J(u_m) - \inf J(\mathcal{H})\big)_{m\in\mathbb{N}}$.

(vi): Since $\big(J(u_m) - \inf J(\mathcal{H})\big)_{m\in\mathbb{N}}$ is summable, it follows from (4.10) and [15, Lemma 3] that $\big(J(u_m) - \inf J(\mathcal{H})\big) = o(1/m)$. □

The purpose of the rest of the section is to show how approximations of the type considered in Theorem 2.7 (equation (2.15)) can be computed explicitly.

Lemma 4.2 *Let $h\colon \mathbb{R}\to\mathbb{R}$ be convex and such that $0 \in \operatorname{Argmin}_{\mathbb{R}} h$, let $(s,\mu)\in\mathbb{R}^2$, and let $\alpha \in [-1,1]$. Let $\beta \in \mathbb{R}_+$ be the Lipschitz constant of h in $[-1-|\mu|, 1+|\mu|]$ and set*

$$\delta = \sqrt{(2\beta + 2|\mu| + 1)|\alpha|} \qquad \text{and} \qquad s = \operatorname{prox}_h \mu + \alpha. \tag{4.14}$$

Then $s \simeq_\delta \operatorname{prox}_h \mu$. Moreover, $\widehat{s} = \operatorname{sign}(\mu)\max\{0, \operatorname{sign}(\mu)s\}$ satisfies $\widehat{s} \simeq_\delta \operatorname{prox}_h \mu$ and $\mu\widehat{s} \geqslant 0$.

Proof. Let $t = \operatorname{prox}_h \mu$. Since $0 \in \operatorname{Argmin} h$, $\operatorname{prox}_h 0 = 0$. Hence, since prox_h is nonexpansive and increasing [12, Lemma 2.4], $|t| \leqslant |\mu|$ and $\operatorname{sign}(t) = \operatorname{sign}(\mu)$. We note that $|s| \leqslant |s-t| + |t| \leqslant 1 + |\mu|$. Thus,

$$\begin{aligned} h(s) + \frac{1}{2}|s-\mu|^2 - h(t) - \frac{1}{2}|t-\mu|^2 &\leqslant \beta|s-t| + \frac{1}{2}|s-t||s-\mu+t-\mu| \\ &\leqslant \frac{1}{2}(2\beta + 1 + 2|\mu|)|\alpha|. \end{aligned} \tag{4.15}$$

To conclude, it is enough to note that $|\widehat{s} - \operatorname{prox}_h(\mu)| \leqslant |\alpha|$. □

Lemma 4.3 *Let $h \in \Gamma_0(\mathbb{R})$, let $\sigma \in \Gamma_0(\mathbb{R})$ be a support function, and set $\phi = h + \sigma$. Let $(s,x)\in\mathbb{R}^2$ be such that $s\operatorname{prox}_\sigma(x) \geqslant 0$, and let $\delta \in \mathbb{R}_+$. Then*

$$s \simeq_\delta \operatorname{prox}_h(\operatorname{prox}_\sigma x) \quad \Rightarrow \quad s \simeq_\delta \operatorname{prox}_\phi x. \tag{4.16}$$

Proof. Let $\mu = \operatorname{prox}_\sigma x$ and $s \simeq_\delta \operatorname{prox}_h(\operatorname{prox}_\sigma x)$. By [27, Lemma 2.4] there exist $(\delta_1, \delta_2) \in \mathbb{R}_+^2$ and $e \in \mathbb{R}$, such that

$$\mu - s + e \in \partial_{\delta_1^2/2} h(s), \quad |e| \leqslant \delta_2, \quad \text{and} \quad \delta_1^2 + \delta_2^2 \leqslant \delta^2. \tag{4.17}$$

Hence

$$x - s + e = x - \mu + \mu - s + e \in \partial\sigma(\mu) + \partial_{\delta_1^2/2} h(s). \tag{4.18}$$

Since $s\mu \geqslant 0$, there exists $t \in \mathbb{R}_+$ such that $\mu = ts$. Moreover, since σ is positively homogeneous, $\partial\sigma(ts) \subset \partial\sigma(s)$. Therefore $x - s + e \in \partial\sigma(s) + \partial_{\delta_1^2/2} h(s) \subset \partial_{\delta_1^2/2}\phi(s)$, which implies that $s \simeq_\delta \operatorname{prox}_\phi x$ by [27, Lemma 2.4]. □

Remark 4.4 Let $h \in \Gamma_0(\mathbb{R})$, let $(s, \mu) \in \mathbb{R}^2$, and let $\delta \in \mathbb{R}_{++}$. Suppose that $0 \in \operatorname{Argmin}_{\mathbb{R}} h$ and that $s \simeq_\delta \operatorname{prox}_h \mu$ with $\delta \leqslant |s|$. Then $s\mu \geqslant 0$. Indeed, since $h(0) = \inf h(\mathbb{R})$, we have

$$h(s) + \frac{1}{2}|s - \mu|^2 \leqslant h(0) + \frac{1}{2}\mu^2 + \frac{1}{2}\delta^2 \leqslant h(s) + \frac{1}{2}\mu^2 + \frac{1}{2}\delta^2 \tag{4.19}$$

and hence $0 \leqslant (1/2)(s^2 - \delta^2) \leqslant s\mu$. This shows that Lemma 4.3, when $\delta = 0$, gives $\operatorname{prox}_\phi = \operatorname{prox}_h \circ \operatorname{prox}_\sigma$ and consequently generalizes [9, Proposition 3.6], relaxing also the condition on the differentiability of h at 0. With the help of this result one can compute general thresholders operators as the proximity operator of $|\cdot| + \eta|\cdot|^r$. Figure 1 depicts some instances of these thresholders (see also [9]).

The following lemma is an error-tolerant version of [10, Proposition 12].

Lemma 4.5 *Let $\phi \in \Gamma_0(\mathbb{R})$, let $(s, x, p) \in \mathbb{R}^3$, let $\delta \in \mathbb{R}_+$, and let $C \subset \mathbb{R}$ be a nonempty closed interval. Then*

$$s \simeq_\delta \operatorname{prox}_\phi x, \quad \textit{and} \quad p = \operatorname{proj}_C s \quad \Rightarrow \quad p \simeq_\delta \operatorname{prox}_{\phi + \iota_C} x. \tag{4.20}$$

Proof. Let $g = \phi + (1/2)(\cdot - x)^2$ and let $\epsilon = (\delta^2/2)$. Since g is convex and $\bar{s} = \operatorname{prox}_\phi x$ is its minimum, g is decreasing on $]-\infty, \bar{s}]$ and increasing on $[\bar{s}, +\infty[$. By definition s is a ϵ-minimizer of g. The statement is equivalent to the fact that p is a ϵ-minimizer of $g + \iota_C$. If $s \in C$, then p is a fortiori an ϵ-minimizer of $g + \iota_C$. We now consider two cases. First suppose that $s < \inf C$. If $s < \inf C \leqslant \bar{s}$, then $\inf C$ is still an ϵ-minimizer of g and $\inf C \in C$. Thus $p = \inf C$ is an ϵ-minimizer of $g + \iota_C$. If either $s \leqslant \bar{s} \leqslant \inf C$ or $\bar{s} \leqslant s < \inf C$, we have $p = \operatorname{proj}_C \bar{s} = \inf C$, which is the minimum of $g + \iota_C$, since g is increasing on $[\bar{s}, +\infty[$. The second case $\sup C < s$ is treated likewise. □

Proposition 4.6 *Let $\mathcal{H}$ be a separable real Hilbert space and let $(o_k)_{k\in\mathbb{K}}$ be an orthonormal basis of $\mathcal{H}$, where $\mathbb{K}$ is an at most countable set. Let $(h_k)_{k\in\mathbb{K}}$ be a family of convex functions from $\mathbb{R}$ to $\mathbb{R}$ such that, for every $k \in \mathbb{K}$, $h_k \geqslant h_k(0) = 0$. Let $(C_k)_{k\in\mathbb{K}}$ be a family of closed intervals in $\mathbb{R}$ such that $0 \in \bigcap_{k\in\mathbb{K}} C_k$, let $(D_k)_{k\in\mathbb{K}}$ be a family of nonempty closed bounded intervals in $\mathbb{R}$. Suppose that $(h_k^*(-(\inf D_k)_+))_{k\in\mathbb{K}}$ and $(h_k^*((\sup D_k)_-))_{k\in\mathbb{K}}$ are summable, and set*

$$G \colon \mathcal{H} \to \,]-\infty, +\infty] : u \mapsto \sum_{k\in\mathbb{K}} (\iota_{C_k} + \sigma_{D_k} + h_k)(\langle u \mid o_k \rangle). \tag{4.21}$$

Let $w \in \mathcal{H}$, let $(\alpha_k)_{k\in\mathbb{K}} \in \mathbb{R}^{\mathbb{K}}$, let $(\xi_k)_{k\in\mathbb{K}} \in \ell^1(\mathbb{K})$, set $\delta = \sqrt{\sum_{k\in\mathbb{K}} \xi_k}$, and let

$$
\begin{array}{l}
\text{for every } k \in \mathbb{K} \\
\left\lfloor
\begin{array}{l}
\chi_k = \langle w \mid o_k \rangle \\
|\alpha_k| \leqslant \dfrac{\xi_k}{4\gamma \max\{h_k(|\chi_k|+2), h_k(-|\chi_k|-2))\} + 2|\chi_k| + 1} \\
\pi_k = \operatorname{prox}_{\gamma h_k}\big(\operatorname{soft}_{\gamma D_k} \chi_k\big) + \alpha_k \\
\nu_k = \operatorname{proj}_{C_k}\big(\operatorname{sign}(\chi_k) \max\big\{0, \operatorname{sign}(\chi_k)\pi_k\big\}\big).
\end{array}
\right.
\end{array}
\tag{4.22}
$$

Now set $v = \sum_{k\in\mathbb{K}} \nu_k o_k$. Then $v \simeq_\delta \operatorname{prox}_{\gamma G} w$.

Proof. The function G lies in $\Gamma_0(\mathcal{H})$ as the composition of the linear isometry $\mathcal{H} \to \ell^2(\mathbb{K})\colon u \mapsto (\langle u \mid o_k \rangle)_{k\in\mathbb{N}}$ and the function

$$
\ell^2(\mathbb{K}) \to \left]-\infty, +\infty\right] : (\mu_k)_{k\in\mathbb{K}} \mapsto \sum_{k\in\mathbb{K}} g_k(\mu_k), \quad \text{with} \quad g_k = \iota_{C_k} + \sigma_{D_k} + h_k,
\tag{4.23}
$$

which belongs to $\Gamma_0(\ell^2(\mathbb{K}))$ by Lemma A.1. Now set

$$
(\forall k \in \mathbb{K}) \quad
\begin{cases}
\mu_k = \operatorname{soft}_{\gamma D_k} \chi_k \\
s_k = \operatorname{sign}(\mu_k) \max\big\{0, \operatorname{sign}(\mu_k)(\operatorname{prox}_{\gamma h_k} \mu_k + \alpha_k)\big\} \\
\nu_k = \operatorname{proj}_{C_k} s_k.
\end{cases}
\tag{4.24}
$$

Let $k \in \mathbb{K}$. Since $\operatorname{soft}_{\gamma D_k}$ is nonexpansive and $2\gamma \max\{h_k(|\chi_k|+2), h_k(-|\chi_k|-2)\}$ is a Lipschitz constant for γh_k on the interval $[-|\chi_k|-1, |\chi_k|+1]$, it follows from (4.24) and Lemma 4.2 that

$$
\begin{cases}
\delta_k^2 = \big(4\gamma \max\{h_k(|\chi_k|+2), h_k(-|\chi_k|-2)\} + 2|\chi_k| + 1\big)|\alpha_k| \\
s_k \simeq_{\delta_k} \operatorname{prox}_{\gamma h_k}(\operatorname{prox}_{\gamma\sigma_{D_k}} \chi_k) \\
s_k \operatorname{prox}_{\gamma\sigma_{D_k}} \chi_k \geqslant 0.
\end{cases}
\tag{4.25}
$$

Thus, Lemma 4.3 yields

$$
s_k \simeq_{\delta_k} \operatorname{prox}_{\gamma(h_k+\sigma_{D_k})} \chi_k,
\tag{4.26}
$$

and, using Lemma 4.5, we obtain $\nu_k \simeq_{\delta_k} \operatorname{prox}_{\gamma g_k} \chi_k$. Hence, by Definition 2.6,

$$
\gamma g_k(\nu_k) + \frac{1}{2}|\nu_k - \chi_k|^2 \leqslant \gamma g_k\big(\operatorname{prox}_{\gamma g_k} \chi_k\big) + \frac{1}{2}|\operatorname{prox}_{\gamma g_k} \chi_k - \chi_k|^2 + \frac{\delta_k^2}{2}.
\tag{4.27}
$$

On the other hand, we derive from [12, Example 2.19] and [9, Proposition 3.6] that

$$
\langle \operatorname{prox}_{\gamma G} w \mid o_k \rangle = \operatorname{prox}_{\gamma g_k} \chi_k.
\tag{4.28}
$$

Thus, summing the inequalities (4.27) over k, we obtain

$$
\begin{aligned}
&\gamma \sum_{k\in\mathbb{K}} g_k(\nu_k) + \frac{1}{2}\sum_{k\in\mathbb{K}} |\nu_k - \chi_k|^2 \\
&\leqslant \gamma \sum_{k\in\mathbb{K}} \Big(g_k\big(\langle \operatorname{prox}_{\gamma g} w \mid o_k \rangle\big) + \frac{1}{2}\big|\langle \operatorname{prox}_{\gamma G} w \mid o_k \rangle - \langle w \mid o_k \rangle\big|^2 \Big) + \frac{1}{2}\sum_{k\in\mathbb{K}} \delta_k^2 \\
&\leqslant \gamma G\big(\operatorname{prox}_{\gamma G} w\big) + \frac{1}{2}\|\operatorname{prox}_{\gamma G} w - w\|^2 + \frac{1}{2}\sum_{k\in\mathbb{K}} \xi_k \\
&< +\infty.
\end{aligned}
\tag{4.29}
$$

Thus, (A.5) and (4.29) yield $(\nu_k)_{k\in\mathbb{N}} \in \ell^2(\mathbb{K})$ and one can find $v \in \mathcal{H}$ such that, for every $k \in \mathbb{K}$, $\chi_k = \nu_k$. Hence,

$$\gamma G(v) + \frac{1}{2}\|v - w\|^2 \leqslant \gamma G(\operatorname{prox}_{\gamma G} w) + \frac{1}{2}\|\operatorname{prox}_{\gamma G} w - w\|^2 + \frac{1}{2}\sum_{k\in\mathbb{K}} \xi_k \tag{4.30}$$

and finally $v \simeq_\delta \operatorname{prox}_{\gamma G} w$, where $\delta = \sqrt{\sum_{k\in\mathbb{K}} \xi_k}$. □

Proof. [of Theorem 2.11] (i): Lemma A.1 guarantees that $G \in \Gamma_0(\ell^2(\mathbb{K}))$, that G is coercive, and that $\operatorname{dom} G \subset \ell^r(\mathbb{K})$. The statement therefore follows from [3, Corollary 11.15(ii)].

(ii)–(iv): Let $\widehat{F}_n \colon \mathcal{H} \to \mathbb{R} \colon u \to (1/n)\sum_{i=1}^n (Au(x_i) - y_i)^2$. Then, for every $u \in \ell^2(\mathbb{K})$, $\nabla \widehat{F}_n(u) = (2/n)\sum_{i=1}^n (\langle u \mid \Phi(x_i)\rangle - y_i)\Phi(x_i)$. Hence, since $\|\Phi(x_i)\|_2 \leqslant \kappa$, $\nabla \widehat{F}_n$ is Lipschitz continuous with constant $2\kappa^2$. Therefore, the statement follows from Theorem 2.7 and Proposition 4.6. It remains to show the convergence properties of $(\|u_m - \widehat{u}\|_r)_{m\in\mathbb{N}}$ and $(\|v_m - \widehat{u}\|_r)_{m\in\mathbb{N}}$. We focus on the sequence $(\|u_m - \widehat{u}\|_r)_{m\in\mathbb{N}}$, since $(\|v_m - \widehat{u}\|_r)_{m\in\mathbb{N}}$ can be treated analogously. It follows from Lemma 3.1 and the convexity of $\widehat{F}_n$ that

$$(\forall\, m \in \mathbb{N}) \quad (\widehat{F}_n + \lambda G)(u_m) - (\widehat{F}_n + \lambda G)(\widehat{u}) \geqslant \frac{\eta\lambda M \|u_m - \widehat{u}\|_r^2}{\big(\|\widehat{u}\|_r + \|u_m - \widehat{u}\|_r\big)^{2-r}}. \tag{4.31}$$

Therefore, since $(\widehat{F}_n + \lambda G)(u_m) - (\widehat{F}_n + \lambda G)(\widehat{u}) \to 0$ as $m \to +\infty$ and $\psi \colon \mathbb{R}_+ \to \mathbb{R} \colon t \mapsto t^2/(\|\widehat{u}\| + t)^{2-r}$ is strictly increasing with $\psi(0) = 0$, we obtain $\|u_m - \widehat{u}\|_r \to 0$. Moreover, taking $\rho \in \mathbb{R}_{++}$ such that $\sup_{m\in\mathbb{N}} \big(\|\widehat{u}\|_r + \|u_m - \widehat{u}\|_r\big)^{2-r} \leqslant \rho$, (2.18) follows from (4.31). □

References

[1] A. Antoniadis, D. Leporini, and J.-C. Pesquet. Wavelet thresholding for some classes of non-Gaussian noise, *Statistica Neerlandica*, 56:434–453, 2002.

[2] H. Attouch. Viscosity solutions of minimization problems, *SIAM Journal on Optimization,* 6:769–805, 1996.

[3] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* Springer, New York, 2011.

[4] K. Bredies. A forward-backward splitting algorithm for the minimization of non-smooth convex functionals in Banach space. *Inverse Problems*, 25: art. 015005, 2009.

[5] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data*. Springer, Heidelberg, 2011.

[6] D. Butnariu, A. N. Iusem, and C. Zălinescu. On uniform convexity, total convexity and convergence of the proximal point and outer Bregman projection algorithms in Banach spaces. *Journal of Convex Analysis*, 10:35–61, 2003.

[7] C. Chaux, P. L. Combettes, J.-C. Pesquet, and V. Wajs. A variational formulation for frame-based inverse problems. *Inverse Problems*, 23:1495–1518, 2007.

[8] P. L. Combettes. Strong convergence of block-iterative outer approximation methods for convex optimization. *SIAM Journal on Control and Optimization*, 38:538–565, 2000.

[9] P. L. Combettes and J.-C. Pesquet. Proximal thresholding algorithm for minimization over orthonormal bases. *SIAM Journal on Optimization*, 18:1351–1376, 2007.

[10] P. L. Combettes and J.-C. Pesquet. A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing*, 1:564–574, 2007.

[11] P. L. Combettes, S. Salzo, and S. Villa. Consistency of regularized learning schemes in Banach spaces. arXiv:1410.6847, 2014.

[12] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, 4:1168–1200, 2005.

[13] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society (N.S.)*, 39:1–49, 2002.

[14] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57:1413–1457, 2004.

[15] D. Davis and Y. Yin. Convergence rate analysis of several splitting schemes. arXiv:1406.4834v3, 2015.

[16] C. De Mol, E. De Vito, and L. Rosasco. Elastic-net regularization in learning theory. *Journal of Complexity*, 25:201–230, 2009.

[17] E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri. Some properties of regularized kernel methods. *Journal of Machine Learning Research*, 5:1363–1390, 2004.

[18] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:883–904, 2005.

[19] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1-50, 2000.

[20] W. J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998.

[21] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

[22] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[23] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 45:7–57, 2009.

[24] J. J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien, *Comptes Rendus de l'Académie des Sciences de Paris,* 255:2897–2899, 1962.

[25] Yu. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming, Ser. B*, 140:125–161, 2013.

[26] B. T. Polyak. *Introduction to Optimization.* Optimization Software Inc., New York, 1987.

[27] S. Salzo and S. Villa. Inexact and accelerated proximal point algorithms. *Journal of Convex Analysis*, 19:1167–1192, 2012.

[28] M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems*, 24:1458–1466, 2011.

[29] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In D. Helmbold and B. Williamson, editors, *Computational Learning Theory*, volume 2111 of *Lecture Notes in Computer Science*, pp. 416–426. Springer, Berlin, 2001.

[30] S. Villa, S. Salzo, L. Baldassarre, and A. Verri. Accelerated and inexact forward-backward algorithms. *SIAM Journal on Optimization*, 23:1607–1633, 2013.

[31] V. Wajs. Décompositions et algorithmes proximaux pour l'analyse et le traitement itératif des signaux, 2007. Thèse de doctorat, Université Pierre et Marie Curie, Paris.

[32] V. Yurinsky. *Sums and Gaussian Vectors*, volume 1617 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1995.

[33] C. Zălinescu. *Convex Analysis in General Vector Spaces*. World Scientific, River Edge, NJ, 2002.

[34] Z. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67:301–320, 2005.

A An auxiliary result

The following result is a generalization of [12, Proposition 5.14].

Lemma A.1 *Let $\mathbb{K}$ be an at most countable set. For every $k \in \mathbb{K}$, let C_k be a closed interval in $\mathbb{R}$ such that $0 \in C_k$, let D_k be a nonempty closed bounded interval in $\mathbb{R}$, and let $h_k \in \Gamma_0^+(\mathbb{R})$ be such that $h_k(0) = 0$. Set*

$$G\colon \ell^2(\mathbb{K}) \to \left]-\infty,+\infty\right] : (\xi_k)_{k\in\mathbb{K}} \mapsto \sum_{k\in\mathbb{K}} g_k(\xi_k), \quad \textit{where} \quad g_k = \iota_{C_k} + \sigma_{D_k} + h_k. \tag{A.1}$$

Let $r \in \left]1,2\right]$ and consider the following statements:

(a) $\sum_{k\in\mathbb{K}} |(\inf D_k)_+|^2 < +\infty$ *and* $\sum_{k\in\mathbb{K}} |(\sup D_k)_-|^2 < +\infty$.

(b) $\sum_{k\in\mathbb{K}} h_k^*\big(-(\inf D_k)_+\big) < +\infty$ *and* $\sum_{k\in\mathbb{K}} h_k^*((\sup D_k)_-) < +\infty$.

(c) $\sum_{k\in\mathbb{K}} |(\inf D_k)_+|^{r^*} < +\infty$ *and* $\sum_{k\in\mathbb{K}} |(\sup D_k)_-|^{r^*} < +\infty$.

Then the following hold:

(i) *Suppose that* (a) *or* (b) *is satisfied. Then* $G \in \Gamma_0(\ell^2(\mathbb{K}))$.

(ii) *Suppose that* (b) *is satisfied. Then* $\inf\, G(\ell^2(\mathbb{K})) > -\infty$.

(iii) *Suppose that, for every* $k \in \mathbb{K}$, $h_k \geqslant \eta|\cdot|^r$ *for some* $\eta \in \mathbb{R}_{++}$. *Then* (a)$\Rightarrow$(c)$\Rightarrow$(b).

(iv) *Suppose that, for every* $k \in \mathbb{K}$, $h_k - \eta|\cdot|^r \in \Gamma_0^+(\mathbb{R})$ *for some* $\eta \in \mathbb{R}_{++}$ *and that* (c) *holds. Then, for every* $\eta' \in \left]0, \eta\right[$, *there exists* $H \in \Gamma_0(\ell^2(\mathbb{K}))$ *such that* $G\colon u \mapsto H(u) + \eta' \sum_{k\in\mathbb{K}} |\mu_k|^r$, $\operatorname{dom} G \subset \ell^r(\mathbb{K})$, *and* G *is coercive in* $\ell^2(\mathbb{K})$.

Proof. We first observe that, if there exist $(\chi_k)_{k\in\mathbb{K}} \in \ell^1_+(\mathbb{K})$ and $b \in \mathbb{R}_+$ such that

$$(\forall k \in \mathbb{K}) \quad -g_k \leqslant \chi_k + b|\cdot|^2, \tag{A.2}$$

then $G \in \Gamma_0(\ell^2(\mathbb{K}))$.

(i): Let $k \in \mathbb{K}$. Since

$$(\forall \mu \in \mathbb{R}) \quad \sigma_{D_k}(\mu) = \begin{cases} \mu \sup D_k & \text{if } \mu \geqslant 0 \\ \mu \inf D_k & \text{if } \mu < 0, \end{cases} \tag{A.3}$$

we have

$$\begin{aligned}(\forall \mu \in \mathbb{R}) \quad -g_k(\mu) &\leqslant -\sigma_{D_k}(\mu) - h_k(\mu) \\ &\leqslant \max\{(\mu)_-(\inf D_k)_+, (\mu)_+(\sup D_k)_-\} - h_k(\mu).\end{aligned} \tag{A.4}$$

Hence, in order to guarantee condition (A.2) for some $(\chi_k)_{k\in\mathbb{K}} \in \ell^1_+(\mathbb{K})$ and $b \in \mathbb{R}_{++}$, it is sufficient to require condition (a) or (b) (note that $h_k^* \geqslant 0$, since $h_k(0) = 0$). Therefore in this case $G \in \Gamma_0(\ell^2(\mathbb{K}))$.

(ii): It follows from (A.4) that

$$(\forall\, k \in \mathbb{K}) \qquad -g_k \leqslant \max\big\{h_k^*\big(-(\inf D_k)_+\big), h_k^*\big((\sup D_k)_-\big)\big\}. \tag{A.5}$$

Hence, for every $u \in \ell^2(\mathbb{K})$, $-G(u) \leqslant \sum_{k\in\mathbb{K}} \max\big\{h_k^*\big(-(\inf D_k)_+\big), h_k^*\big((\sup D_k)_-\big)\big\} < +\infty$.

(iii): For every $k \in \mathbb{K}$, $h_k^* \leqslant (r\eta)^{1-r^*}(r^*)^{-1}|\cdot|^{r^*}$. The statement therefore follows by observing that, since $2 \leqslant r^*$, $\ell^2(\mathbb{K}) \subset \ell^{r^*}(\mathbb{K})$.

(iv): Setting, for every $k \in \mathbb{K}$, $\tilde{h}_k = h_k - \eta'|\cdot|^r$, we have $g_k = \iota_{C_k} + \sigma_{D_k} + \tilde{h}_k + \eta'|\cdot|^r$, with $(\eta - \eta')|\cdot|^r \leqslant \tilde{h}_k \in \Gamma_0^+(\mathbb{R})$. It follows from (i) and (iii) that, for every $u = (\mu_k)_{k\in\mathbb{K}} \in \ell^2(\mathbb{K})$, $G(u) = H(u) + \eta' \sum_{k\in\mathbb{K}} |\mu_k|^r$, for some $H \in \Gamma_0(\ell^2(\mathbb{K}))$. ☐

B Proximity operators of power functions

It follows from [7, Example 4.4] that, for every $\gamma \in \mathbb{R}_{++}$ and every $r \in [1, 2]$,

$$(\forall\, \mu \in \mathbb{R}) \quad \operatorname{prox}_{\gamma|\cdot|^r} \mu = \xi \operatorname{sign}(\mu), \quad \text{where} \quad \xi \geqslant 0 \quad \text{and} \quad \xi + r\gamma\xi^{r-1} = |\mu|. \tag{B.1}$$

There are several exponents r for which Equation (B.1) can be solved explicitly for $r \in \{3/2, 4/3, 5/4\}$ [7, 31]. However, in general, it must be solved iteratively.

Proposition B.1 *Let $\mu \in \mathbb{R}$, let $\gamma \in \mathbb{R}_{++}$, let $r \in [1,2]$, and let $(r_1, r_2) \in [1,2]^2$, be such that $r_1 < r_2$. Then the following hold:*

(i) $\operatorname{prox}_{\gamma|\cdot|^r} \colon \mathbb{R} \to \mathbb{R}$ *is strictly increasing, nonexpansive, odd, and differentiable, and* $\operatorname{prox}_{\gamma|\cdot|^r} + \iota_{\mathbb{R}_+}$ *is convex.*

(ii) *We have*

$$\min\left\{\frac{|\mu|}{1+r\gamma}, \left(\frac{|\mu|}{1+r\gamma}\right)^{\frac{1}{r-1}}\right\} \leqslant |\operatorname{prox}_{\gamma|\cdot|^r} \mu| \leqslant \max\left\{\frac{|\mu|}{1+r\gamma}, \left(\frac{|\mu|}{1+r\gamma}\right)^{\frac{1}{r-1}}\right\}. \tag{B.2}$$

(iii) *Suppose that $|\mu| > 1 + r_2\gamma$. Then $|\operatorname{prox}_{\gamma|\cdot|^{r_2}} \mu| < |\operatorname{prox}_{\gamma|\cdot|^{r_1}} \mu|$.*

(iv) *Suppose that $r > 1$ and that $|\mu| > 1 + r\gamma$. Then $\dfrac{|\mu|}{1+r\gamma} \leqslant |\operatorname{prox}_{\gamma|\cdot|^r} \mu| < |\mu| - \gamma$.*

Proof. (i): It follows from [9, Lemma 2.2(iv) and Proposition 2.4] that $\operatorname{prox}_{\tau|\cdot|}$ is nonexpansive, increasing, and odd. Now set $\psi\colon \mathbb{R}_+ \to \mathbb{R}_+ \colon \xi \mapsto \xi + r\tau\xi^{r-1}$. Clearly ψ is strictly increasing and concave. Moreover it is differentiable on $\mathbb{R}_{++}$ and, for every $\xi \in \mathbb{R}_{++}$, $\psi'(\xi) = 1 + r(r-1)/\xi^{2-r}$. Hence, from (B.1), for every $\mu \in \mathbb{R}_+$, $\operatorname{prox}_{\tau|\cdot|^r} \mu = \psi^{-1}(\mu)$. This shows that $\operatorname{prox}_{\tau|\cdot|^r}$ is strictly increasing, convex, differentiable on $\mathbb{R}_{++}$ with, for every $\mu \in \mathbb{R}_{++}$, $(\operatorname{prox}_{\tau|\cdot|^r})'\mu = 1/\psi'(\psi^{-1}(\mu))$, that is

$$(\operatorname{prox}_{\gamma|\cdot|^r})'\mu = \left(1 + \frac{r(r-1)\gamma}{(\operatorname{prox}_{\gamma|\cdot|^r} \mu)^{2-r}}\right)^{-1}. \tag{B.3}$$

(ii): According to (B.1), there exists $\xi \in \mathbb{R}_+$ such that $\operatorname{prox}_{\tau|\cdot|^r} \mu = \operatorname{sign}(\mu)\xi$ and $\xi + r\tau\xi^{r-1} = |\mu|$. If $\xi \geqslant 1$, then $|\mu| = \xi + r\tau\xi^{r-1} \leqslant (1+r\tau)\xi$, hence $|\mu|/(1+r\tau) \leqslant \xi = |\operatorname{prox}_{\tau|\cdot|^r} \mu|$. If $\xi < 1$, then $|\mu| = \xi + r\tau\xi^{r-1} \leqslant (1+r\tau)\xi^{r-1}$, hence $\big(|\mu|/(1+r\tau)\big)^{1/(r-1)} \leqslant \xi = |\operatorname{prox}_{\tau|\cdot|^r} \mu|$. The first inequality in (B.2) follows and the second is proved analogously.

(iii): In view of (B.1) there exist $\xi_1 \in \mathbb{R}_+$ and $\xi_2 \in \mathbb{R}_+$ such that

$$\begin{cases} \operatorname{prox}_{\tau|\cdot|^{r_1}} \mu = \operatorname{sign}(\mu)\xi_1 \quad \text{and} \quad \xi_1 + r_1\tau\xi_1^{r_1-1} = |\mu| \\ \operatorname{prox}_{\tau|\cdot|^{r_2}} \mu = \operatorname{sign}(\mu)\xi_2 \quad \text{and} \quad \xi_2 + r_2\tau\xi_2^{r_2-1} = |\mu|. \end{cases} \tag{B.4}$$

If $|\mu| > 1 + \tau r_2 > 1 + \tau r_1$, it follows from (B.2) that

$$1 < \frac{|\mu|}{1+r_1\tau} \leqslant |\xi_1| \quad \text{and} \quad 1 < \frac{|\mu|}{1+r_2\tau} \leqslant |\xi_2|. \tag{B.5}$$

Therefore, since $r_1 < r_2$ and $\xi_1 > 1$,

$$\xi_2 + r_2\tau\xi_2^{r_2-1} = |\mu| = \xi_1 + r_1\tau\xi_1^{r_1-1} < \xi_1 + r_2\tau\xi_1^{r_2-1}. \tag{B.6}$$

Hence, since $\xi \mapsto \xi + r_2\tau\xi^{r_2-1}$ is strictly increasing on $\mathbb{R}_+$, we conclude that $\xi_2 < \xi_1$.

(iv): Since (B.1) implies that $\operatorname{prox}_{\tau|\cdot|} \mu = \operatorname{sign}(\mu)(|\mu| - \tau)$, we derive from (iii) that

$$|\mu| > 1 + r\tau \quad \Rightarrow \quad |\operatorname{prox}_{\tau|\cdot|^r} \mu| < |\mu| - \tau, \tag{B.7}$$

The first inequality in (iv) follows directly from (B.2). ☐

Remark B.2

(i) The bounds given in (B.2) can be useful to initialize the bisection method to solve (B.1).

(ii) $(\operatorname{prox}_{\gamma|\cdot|^r})'0 = 0$, $(\operatorname{prox}_{\gamma|\cdot|^r})'\mu \leqslant 1$ and $(\operatorname{prox}_{\gamma|\cdot|^r})'\mu \to 1$ as $\mu \to +\infty$.

(iii) $\operatorname{prox}_{\gamma|\cdot|^r}$ has no asymptote as $\mu \to +\infty$, since (B.1) yields $\operatorname{prox}_{\gamma|\cdot|^r} \mu - \mu = -r\gamma(\operatorname{prox}_{\gamma|\cdot|^r} \mu)^{r-1} \to -\infty$ as $\mu \to +\infty$.